



(19)

(11) Publication number:

0

Generated Document.

PATENT ABSTRACTS OF JAPAN

(21) Application number: **05051663**(51) Int'l. Cl.: **G06F 3/06 G06F 3/06 G06F**(22) Application date: **12.03.93**

(30) Priority:

(43) Date of application
publication: **22.09.94**(84) Designated contracting
states:(71) Applicant: **HITACHI LTD**(72) Inventor: **TSUNODA HITOSHI
TAKAMOTO YOSHIFU
KAMO YOSHIHISA**

(74) Representative:

(54) DISK ARRAY SYSTEM AND DATA WRITE METHOD AND FAULT RECOVERY METHOD FOR THIS SYSTEM

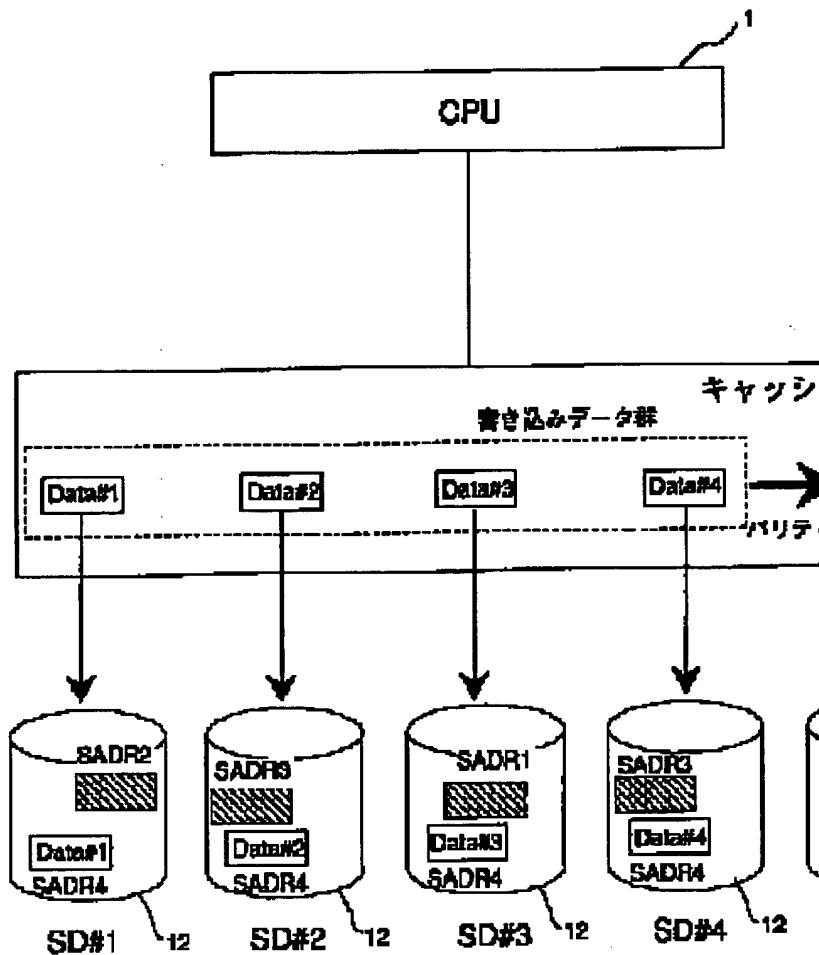
(57) Abstract:

PURPOSE: To reduce the overhead for write with respect to a disk array of RAID (level 5) where data is distributed to improve the processing performance.

CONSTITUTION: Even if data #1 to #4 already written in addresses SADR1 to SADR3 in a drive as data belonging to groups different from one another will be rewritten with write data, these write data are regarded as new write data and are written in the idle area of an address SADR4 in the drive in parallel. Updated old data is not read out. A nullity flag is registered in an address conversion table with respect to updated old data, and data is read from the newly written area. When all of data in original parity groups are

made ineffective, areas holding these groups are used as idle areas.
Effective data in parity groups which are made partially ineffective are justified at a proper timing.

COPYRIGHT: (C)1994,JPO&Japio



(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平6-266510

(43)公開日 平成6年(1994)9月22日

(51)Int.Cl.⁵ 識別記号 廈内整理番号 F I 技術表示箇所
G 0 6 F 3/06 3 0 5 C 7165-5B
3 0 1 Z 7165-5B
3 0 2 A 7165-5B

審査請求 未請求 請求項の数28 O L (全 26 頁)

(21)出願番号	特願平5-51663	(71)出願人	000005108 株式会社日立製作所 東京都千代田区神田駿河台四丁目6番地
(22)出願日	平成5年(1993)3月12日	(72)発明者	角田 仁 東京都国分寺市東恋ヶ窪1丁目280番地 株式会社日立製作所中央研究所内
		(72)発明者	高本 良史 東京都国分寺市東恋ヶ窪1丁目280番地 株式会社日立製作所中央研究所内
		(72)発明者	加茂 善久 東京都国分寺市東恋ヶ窪1丁目280番地 株式会社日立製作所中央研究所内
		(74)代理人	弁理士 小川 勝男

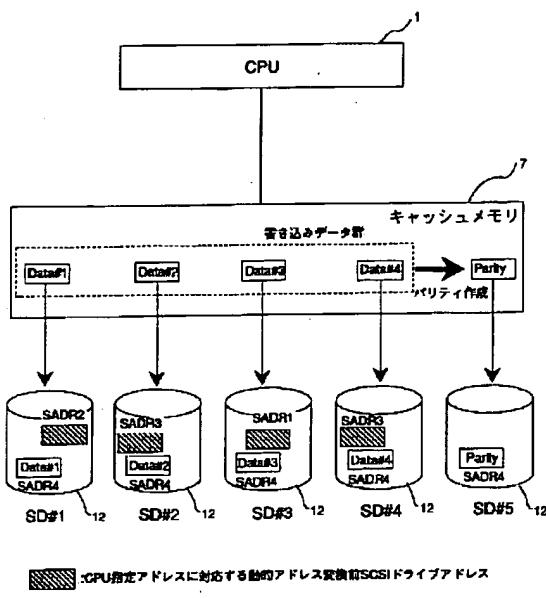
(54)【発明の名称】 ディスクアレイシステムおよびそのためのデータ書き込み方法、障害回復方法

(57)【要約】 (修正有)

【目的】 データを分散させて処理性能を向上させるRAID (LEVEL 5) のディスクアレイにおいて、書き込み時のオーバヘッドを減少させる。

【構成】 書込みデータが、すでにドライブ内アドレスSADR2～SADR3にそれぞれ異なるグループに属するデータとして書き込み済みのData#1#4を書換えるものであっても、新書き込みデータと見なして、ドライブ内のアドレスSADR4の空領域に並列書込む。更新された旧データは読み出さない。更新された旧データに関して無効フラグをアドレス変換テーブルに登録し新に書き込まれた領域から行なう。元のパリティグループ内のすべてのデータが無効にされた場合、そのグループを保持していた領域は、空領域として使用する。部分的に無効にされたパリティグループ内の有効データは、適当なタイミングで詰め替える。

図6



1

2

【特許請求の範囲】

【請求項1】複数のディスクドライブと、一つまたは複数の上位装置から転送された、それぞれ該複数のディスクドライブに新に書き込まれるべき所定の長さのデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定長を有するデータのいずれかからなる複数の書き込みデータを一時的に保持するキャッシュメモリとを有するディスクアレイシステムにおいて、

(a) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正用データグループを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成し、

(b) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブ内の互いに異なるものに属する複数の空き領域に並列に書き込み、

(c) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にするステップからなるデータ書き込み方法。

【請求項2】(d) 該一つまたは複数の上位装置から転送された書き込みデータの量が、該レコード長を有する該所定数に等しい数の書き込みデータを取り出すのに十分になる毎に、上記ステップ(a)から(c)を実行するステップからなるデータ書き込み方法。

【請求項3】該誤り訂正データの生成は、

該一つの上位装置から、一つの書き込みデータが転送されたとき、その一つの書き込みデータをキャッシュ記憶に保持し、

該一つの上位装置から、該一つの書き込みデータが転送されたときに、該キャッシュに既に保持された他の書き込みデータの量が、該一定長を有する該所定数に等しい数の書き込みデータを取り出すのに十分であるときには、該一つまたは複数の上位装置から他の書き込みデータがそのキャッシュに転送されるのを待たないで、上記該誤り訂正データの生成を実行し、

該一つの上位装置から、該一つの書き込みデータが転送されたときに、該キャッシュに保持された書き込みデータの量が、該一定長を有する該所定数に等しい数の書き込みデータを取り出すのに十分でないときには、該キャッシュに保持された書き込みデータの数が、該所定数に達した後、上記該誤り訂正データの生成を実行するステップからなる請求項1記載のデータ書き込み方法。

【請求項4】(d) 該ステップ(a)から(c)を、該キャッシュメモリに保持された、他の書き込みデータに関して繰り返し、

(e) 該ステップ(d)を繰り返し実行した結果、該複数の

10

20

30

40

ドライブに記憶されていたいずれか一つの誤り訂正データグループに属する一群の書き込みデータがすべて無効となったとき、その後のステップ(d)によるステップ(a)から(c)の繰り返し特に、その一群の書き込みデータおよびその誤り訂正データグループに属する誤り訂正用データを保持していた、該複数のドライブ内の複数の領域を、空き領域としていずれか他の誤り訂正データグループの書き込みに使用するステップをさらに有する請求項1記載のデータ書き込み方法。

【請求項5】複数のディスクドライブと、一つまたは複数の上位装置から転送された、それぞれ該複数のディスクドライブに新に書き込まれるべき所定の長さを有するデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定長を有するデータのいずれかからなる複数の書き込みデータを一時的に保持するキャッシュメモリとを有するディスクアレイシステムにおいて、

(a) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正用データグループを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成し、

(b) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブ内の互いに異なるものに属する複数の空き領域に並列に書き込み、

(c) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にするステップからなるデータ書き込み方法。

(d) 該ステップ(a)から(c)を、該キャッシュメモリに保持された、他の書き込みデータに関して繰り返し、

(e) 該複数のドライブの一つに障害が発生したとき、その一つのドライブに保持されていた複数のデータのうち、複数の有効なデータの各々を、該一つのドライブ以外のドライブに保持されている、そのデータが属する誤り訂正データグループに属する、複数の他の有効あるいは無効な書き込みデータと誤り訂正用データとを用いて選択的に回復するステップからなる障害回復方法。

【請求項6】該障害の発生前に、いずれかの一群の書き込みデータに対して、該ステップ(c)を実行した結果、そのステップ(c)で無効なデータとされたいずれか一つのデータが属する誤り訂正データグループに属する一群の書き込みデータがすべて無効なデータとなった場合には、その一群の書き込みデータおよびその誤り訂正グループに属する誤り訂正用データを保持していた、該複数のドライブ内の複数の領域を、空き領域として他の誤り訂正データグループに属するデータの書き込みに使用するステップをさらに有する請求項5記載の障害回復方

法。

【請求項7】該回復ステップは、

該複数のドライブの内、該一つのドライブ以外のドライブより、該複数の他の有効なまたは無効な書き込みデータと該誤り訂正用データとを、選択的に読み出し、該読み出された書き込み済みデータと該読み出された誤り訂正用データとから、該一つの有効なデータを回復するステップからなる請求項5記載の障害回復方法。

【請求項8】複数のディスクドライブと、一つまたは複数の上位装置から転送された、それぞれ該複数のディスクドライブに新に書き込まれるべき所定の長さのデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定長を有するデータのいずれかからなる複数の書き込みデータを一時的に保持するキャッシュメモリとを有するディスクアレイシステムにおいて、

(a) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正用データを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成し、

(b) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブの内の互いに異なるものに属する複数の空き領域に並列に書き込み、

(c) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にするステップからなるデータ書き込み方法。

(d) 該ステップ(a)から(c)を、該キャッシュメモリに保持された、他の書き込みデータに関して繰り返し、

(e) それぞれ少なくとも一つの有効な書き込みデータと少なくとも一つの無効な書き込みデータを含む複数の誤り訂正データグループ(部分的に無効な誤り訂正データグループ)に含まれる複数の有効な書き込みデータを詰め替えるステップからなり、その詰め替えは、(e1)部分的に無効な誤り訂正データグループに含まれている一つまたは複数の有効な書き込みデータを、該複数のドライブから選択的に読み出し、読み出された一つまたは複数の有効な書き込みデータをキャッシュ記憶の保持し、(e2) 該ステップ(e1)を他の部分的に無効な誤り訂正データグループに対して実行し、(e3) 該キャッシュ記憶に保持された、複数の有効な書き込みデータから選択された該所定数に等しい数の有効な書き込みデータの新たな組みとその書き込みデータの組みに対して新に誤り訂正用データを生成し、もって、新たな誤り訂正データグループを生成し、(e4) 該生成された新たな誤り訂正データグループを該複数のドライブ内の空き領域に並列に再度書き込み、(e5) 該ステップ(e3)(e4)を他の新たな

誤り訂正データグループのために繰り返し、(e6) 複数の部分無効誤り訂正データグループであって、そこに含まれていた有効な書き込みデータが該ステップ(e4)(e5)により、すべて再書き込みされたものを保持していた複数の領域を、空き領域として他の誤り訂正データグループの書き込みに使用するステップを有するデータ書き込み方法。

【請求項9】上記ステップ(e1)または(e2)は、特定状態にある複数の部分的に無効な誤り訂正データグループに対して選択的に行なわれる請求項8記載のデータ書き込み方法。

【請求項10】該特定の状態は、部分的に無効な誤り訂正データグループに含まれた無効とされた書き込みデータの数が、予め定めた限界数以上である状態である請求項8記載のデータ書き込み方法。

【請求項11】上記ステップ(e1)または(e2)は、いずれか一つの誤り訂正データグループが該特定状態に達するごとに、その誤り訂正データグループに対して行なわれる請求項9記載のデータ書き込み方法。

【請求項12】上記ステップ(e1)または(e2)は、いずれか一つの該特定状態にある誤り訂正データグループに属する有効な書き込みデータ対していずれか一つの上位装置から転送された読み出し要求に応答してその要求された書き込みデータに対してなされる請求項9記載のデータ書き込み方法。

【請求項13】上記ステップ(e1)または(e2)は、複数の部分的に無効な誤り訂正データグループの数が、所定数に達したときに、該複数の部分的に無効な誤り訂正データグループに対して行なわれる請求項10記載のデータ書き込み方法。

【請求項14】上記ステップ(e1)または(e2)は、該複数のドライブに含まれたデータ書き込み用の領域の内の空き領域の容量が、予め定めた限界値に達したときに、複数の部分的に無効な誤り訂正データグループに対して選択的に行なわれる請求項9記載のデータ書き込み方法。

【請求項15】上記ステップ(e1)または(e2)は、該複数のドライブに含まれたデータ書き込み用の領域の容量に対する空き領域の容量の比率が、予め定めた限界比率に達したときに、複数の部分的に無効な誤り訂正データグループに対して選択的に行なわれる請求項9記載のデータ書き込み方法。

【請求項16】上記一つまたは複数の上位装置から新たな書き込みデータが転送されたとき、該書き込みデータを該キャッシュ記憶に書き込むステップをさらに有し、上記ステップ(e3)は、該キャッシュにいずれかの上位装置から転送された新たな書き込みデータがあるときは、その新たな書き込みデータと該ステップ(e1)で読み出された書き込みデータとからなる該所定数に等しい数の書き込みデータを含む誤り訂正データグループを生成

するステップをさらに有する請求項9記載のデータ書き込み方法。

【請求項17】複数のディスクドライブと、一つまたは複数の上位装置から転送された、それぞれ該複数のディスクドライブに新に書き込まれるべき所定の長さのデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定の長さを有するデータのいずれかからなる複数の書き込みデータを一時的に保持するキャッシュメモリとを有するディスクアレイシステムにおいて、

(a) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正を用データグループを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成し、

(b) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブの内の互いに異なるものに属する複数の空き領域に並列に書き込み、

(c) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にするステップからなるデータ書き込み方法。

(d) 該ステップ(a)から(c)を、該キャッシュメモリに保持された、他の書き込みデータに関して繰り返し、

(e) それぞれ少なくとも一つの有効な書き込みデータと少なくとも一つの無効な書き込みデータを含む誤り訂正データグループ(部分的に無効な誤り訂正データグループ)に含まれる複数の有効な書き込みデータを詰め替えるステップからなり、その詰め替えは、(e1) いずれか一つの上位装置から転送された、いずれか一つの部分的に無効な誤り訂正データグループに含まれている有効な書き込みデータに対する読み出し要求に応答して、その書き込みデータを該複数のドライブから選択的に読み出し、かつ、該読み出された書き込みデータを、該一つの上位装置に転送するとともに、キャッシュ記憶に保持し、(e2) 該ステップ(e1)および(e2)を他の部分的に無効な誤り訂正データグループに対して実行し、(e3) 該キャッシュ記憶に保持された、複数の書き込みデータから選択された該所定数に等しい数の書き込みデータの新たな組みとその書き込みデータの組みに対して新に誤り訂正用データを生成し、もって、新たな誤り訂正データグループを生成し、(e4) 該生成された新たな誤り訂正データグループを該複数のドライブ内の空き領域に並列に書き込み、(e5) 該ステップ(e1)または(e2)により読み出された書き込みデータが、該ステップ(e4)により再書き込みされたとき、該複数のドライブに保持されている、その再書き込みされる前のその書き込みデータを無効にするステップを有するデータ書き込み方法。

【請求項18】上記ステップ(e3)から(e4)は、該ステップ(e1)または(e2)で読み出された書き込みデータが属する部分的に無効な誤り訂正データグループが特定状態にあるときに行なわれる請求項17記載のデータ書き込み方法。

【請求項19】該特定の状態は、部分的に無効な誤り訂正データグループに含まれた無効とされた書き込みデータの数が、予め定めた限界数以上である状態である請求項18記載のデータ書き込み方法。

【請求項20】複数のディスクドライブと、一つまたは複数の上位装置から転送された、それぞれ該複数のディスクドライブに新に書き込まれるべき所定長を有するデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定長のデータのいずれかからなる複数の書き込みデータを一時的に保持するキャッシュメモリとを有するディスクアレイシステムにおいて、

(a) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正を用データグループを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成し、

(b) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブの内の互いに異なるものに属する複数の空き領域に並列に書き込み、

(c) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にするステップからなるデータ書き込み方法。

(d) 該ステップ(a)から(c)を、該キャッシュメモリに保持された、他の書き込みデータに関して繰り返し、

(e) それぞれ少なくとも一つの有効な書き込みデータと少なくとも一つの無効な書き込みデータを含む誤り訂正データグループ(部分的に無効な誤り訂正データグループ)に含まれる複数の有効な書き込みデータを詰め替えるステップからなり、その詰め替えは、(e1) いずれか一つの部分的に無効な誤り訂正データグループに含まれている一つまたは複数の有効な書き込みデータの少なくとも一つを該複数のドライブから選択的に読み出し、かつ、該読み出された書き込みデータを該キャッシュ記憶に一時的に保持し、(e2) いずれか一つの上位装置から新たな書き込みデータが転送してきたときに、その書き込みデータを一時的に該キャッシュ記憶に保持し、(e3)

該ステップ(e1)または(e2)により該キャッシュ記憶に保持された、複数の書き込みデータから該所定数に等しい数の書き込みデータの新たな組みを選択し、その書き込みデータの組みに対して新に誤り訂正用データを生成し、もって、新たな誤り訂正データグループを生成し、(e4) 該生成された新たな誤り訂正データグループを該

複数のドライブ内の空き領域に並列に書き込み、(e5)該ステップ(e1)により読み出された書き込みデータが、該ステップ(e4)により再書き込みされたとき、該複数のドライブに保持されている、その再書き込みされる前のその書き込みデータを無効にするステップを有するデータ書き込み方法。

【請求項21】上記ステップ(e1)は、いずれか一つの部分的に無効な誤り訂正データグループに含まれる有効なデータをすべて読み出すステップからなる請求項20記載のデータ書き込み方法。

【請求項22】上記ステップ(e3)から(e4)は、該ステップ(e1)または(e2)で読み出された書き込みデータが属する部分的に無効な誤り訂正データグループが特定状態にあるときに行なわれる請求項21記載のデータ書き込み方法。

【請求項23】該特定の状態は、部分的に無効な誤り訂正データグループに含まれた無効とされた書き込みデータの数が、予め定めた限界数以上である状態である請求項22記載のデータ書き込み方法。

【請求項24】データの書き込み領域をユーザの要求に従い確保するときに、その領域が動的アドレス変換をする領域か否かをユーザの指示により決定し、

その後上記一つまたは複数の上位装置から上記領域に書き込むべきデータが転送されたときに、上記書き込み領域が動的アドレス変換をする領域と決定されている場合に、上記ステップ(a)から(d)を実行する請求項1記載のデータ書き込み方法。

【請求項25】複数のディスクドライブを有するディスクアレイシステムであって、

(a) それぞれ該複数のディスクドライブに新に書き込まれるべき所定の長さを有するデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定長を有するデータのいずれか一つからなり、一つまたは複数の上位装置から転送された複数の書き込みデータを保持するキャッシュメモリと、

(b) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正用データグループを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成する回路と、

(c) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブの内の互いに異なるものに属する複数の空き領域に並列に書き込む手段と、

(d) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にする手段とを有するもの。

【請求項26】該書き込み手段は、該無効手段による無

効を複数群の書き込みデータに対して実行した結果、該複数のドライブに記憶されていたいはずか一つの誤り訂正データグループに属する一群の書き込みデータがすべて無効となったとき、その一群の書き込みデータおよびその誤り訂正データグループに属する誤り訂正用データを保持していた、該複数のドライブ内の複数の領域を、空き領域として他の誤り訂正データグループの書き込みに使用する手段を有する請求項25記載のディスクアレイシステム

【請求項27】(e) 該複数のドライブの一つに障害が発生したとき、その一つのドライブに保持されていた複数のデータのうち、複数の有効なデータの各々を、該一つのドライブ以外のドライブに保持されている、そのデータが属する誤り訂正データグループに属する、複数の他の有効なあるいは無効な書き込みデータと誤り訂正用データとを用いて選択的に回復する手段をさらに有する請求項25記載のディスクアレイシステム。

【請求項28】(e) それぞれ少なくとも一つの有効な書き込みデータと少なくとも一つの無効な書き込みデータを含む複数の誤り訂正データグループ(部分的に無効な誤り訂正データグループ)に含まれる複数の有効な書き込みデータを詰め替える手段をさらに有し、その詰め替え手段は、(e1) 部分的に無効な誤り訂正データグループに含まれている一つまたは複数の有効な書き込みデータを、該複数のドライブから選択的に読み出し、読み出された一つまたは複数の有効な書き込みデータを該キャッシュメモリに書き込む手段と、(e2) 該書き込み手段(e1)による書き込みを他の部分的に無効な誤り訂正データグループに対して実行する手段と、(e3) 該キャッシュ記憶に保持された、複数の有効な書き込みデータから選択された該所定数に等しい数の有効な書き込みデータの新たな組みとその書き込みデータの組みに対して新に誤り訂正用データを生成し、もって、新たな誤り訂正データグループを生成する手段と、(e4) 該生成された新たな誤り訂正データグループを該複数のドライブ内の空き領域に並列に再度書き込みする手段と、(e5) 該手段(e3)(e4)を他の新たな誤り訂正データグループのために繰返す手段と、(e6) 複数の部分無効誤り訂正データグループであって、そこに含まれていた有効な書き込みデータが該ステップ(d4)(d5)により、すべて再書き込みされたものを保持していた複数の領域を、空き領域として他の誤り訂正データグループの書き込みに使用する手段とを有する請求項25記載のディスクアレイシステム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は高性能な入出力動作を可能とするディスクアレイ装置およびそのためのデータ方法。

【0002】

【従来の技術】現在のコンピュータシステムにおいて

9

は、CPU等の上位側が必要とするデータは2次記憶装置に格納され、CPUが必要とする時に応じ2次記憶装置に対してデータの書き込み、読みだしを行っている。この2次記憶装置としては一般に不揮発な記憶媒体が使用され、代表的なものとして磁気ディスク装置、光ディスクなどがあげられる。以下、ディスク装置をドライブと呼ぶ。

【0003】近年高度情報化に伴い、コンピュータシステムにおいて、2次記憶装置の高性能化が要求されてきた。その一つの解として、多数の比較的容量の小さなドライブにより構成されるディスクアレイが考えられている。

【0004】D.Patterson,G.Gibson, and R.H.Kartz; A Case for Redundant Arrays of Inexpensive Disks(RAID)において、データを分割して並列に処理を行うディスクアレイ(レベル3)とデータを分散して、独立に扱うディスクアレイ(レベル5)について、その性能および信頼性の検討結果が報告されている。

【0005】まず、データを分割して並列に処理を行うレベル3のディスクアレイについて説明する。ディスクアレイは多数の比較的小容量の小さなドライブにより構成される。CPUから転送されてきた一つの書き込みデータを分割し、分割後の複数のデータからパリティデータを作成し、これらのデータを複数のドライブに並列に格納する。また、読み出す場合は逆に各々のドライブから分割されたデータを並列に読み込んできてそれらを結合してCPUへ転送する。複数のデータと誤り訂正用のデータとの群をパリティグループと呼ばれることがある。本明細書でもこの用語を誤り訂正用データがパリティデータでない場合にも用いる。このパリティデータは分割後のデータを格納したドライブの中の1台に障害が発生し、データが読み出せなくなった場合、残りの正常なドライブ内のデータとパリティデータから、障害が発生したドライブ内のデータを復元するためのものである。ディスクアレイのような多数のドライブにより構成される装置では、部品点数が増加することにより、障害が発生する確率が高くなるため、信頼性の向上を図る目的で、このようにパリティデータを用意する。

【0006】次に、データを分散して、独立に扱うレベル5のディスクアレイについて説明する。これは個々のデータを分割せずに独立に扱い、複数のデータからパーティデータを生成し、これらを多数の比較的容量の小さなドライブに分散して格納するものである。このパーティデータも先に述べたように、データを格納したドライブに障害が発生した場合、その障害ディスク内のデータを復元するためのものである。

【0007】現在、一般に使用されている汎用大型コンピュータシステムの2次記憶装置では、他の読み出し/書き込み要求に当該ドライブが使用されているため、そのドライブを使用できずに待たされることが多く発生し

た。このディスクアレイではデータを分散して格納してあるため、読み出し／書き込み要求が増加してもディスクアレイの複数のドライブに分散して処理するため、読み出し／書き込み要求がまたされることが減少する。

【0008】これらのディスクアレイでは、現在一般に使用されている汎用大型コンピュータシステムと同様、2次記憶装置内では、個々のデータの格納場所（アドレス）は予め指定したアドレスに固定され、CPUから当該データへ読みだしおよび書き込みする場合、この固定されたアドレスへアクセスすることになっている。

【0009】データを分散して、独立に扱うレベル5タイプのディスクアレイでは、書き込み時に新しくパーティデータを作成する際に大きな処理オーバヘッドが必要になる。以下それについて説明する。

【0010】図11は上記文献で示したD.Pattersonらが提案したRAIDに述べられている、データを分散して、独立に扱うレベル5タイプのディスクアレイにおける、書き込み時のパーティティデータ作成方法を示す。この各アドレスにあるデータは1回の読み出し／書き込み処理でアクセスされる単位で、個々のデータは独立である。

また、RAIDで述べられているアーキテクチャではデータに対するアドレスは固定されている。前述したようにこのようなシステムでは、信頼性を向上するためパリティデータを設定することが不可欠である。本システムでは各ドライブ内の同一アドレスのデータによりパリティデータが作成される。すなわち、ドライブ#1から4までのアドレス(1..1)のデータによりパリティデータが作成され、パリティデータを格納するドライブの(1..1)に格納される。本システムでは読み出し/書き込み処理は現在の汎用大型計算機システムと同様に各ドライブに対し当該データをアクセスする。このようなディスクアレイにおいて、例えばドライブ#3のアドレス(2..2)にデータを書き込む場合、まず、ドライブ#3の(2..2)の旧データと、ドライブ#5の(2..2)の旧パリティデータを読みだし(ステップ1)、これらと書き込むデータとで排他的論理和をとり、新たなパリティデータを作成する(ステップ2)。

【0011】バリティデータの作成完了後、書き込みデータをドライブ#3の(2.2)に、新バリティデータをドライブ#5の(2.2)に格納する(ステップ3)。

【0012】図12に示すように、レベル5のディスクアレイでは、データの格納されているドライブ、パリティデータの格納されているドライブから古いデータとパリティデータを読み出す時に、ディスクを平均1/2回転待ち、それから読みだしてパリティデータを作成し、この新しく作成したパリティデータとデータを格納するため、さらに一回転待たなければならぬため、データを書き替える場合平均で1.5回転待たなければならぬ。ドライブにおいては1.5回転ディスクの回転を待

11

つということは非常に大きなオーバヘッドとなる。このような書き込み時のオーバヘッドを削減するため、書き込み先のアドレスを動的に変換する方法がストレージテクノロジコーポレーション(以下、STK)社から出願されているPCT国際公開公報WO 91/16711, WO 91/20076に開示されている。

【0013】この従来技術によれば、いずれかの、CPUが指定したデータが更新されるときには、そのデータを含む仮想トラックに属するデータ全体をキャッシュメモリに読み出し、その一部を更新データで更新した上で、更新後の、仮想トラックに属するデータがキャッシュから追い出されたときに、そのデータを物理トラックのセクタ単位に分割し、分割後のデータからあるいはそれらと他の書き込みデータとから新たなパリティグループを生成し、それをドライブ内の空き領域に書き込むようになっている。この際の仮想トラックに属していくデータは無効にされる。パリティグループを構成するデータの長さは、ドライブの一つの物理トラックの容量を有するように定められている。適当なタイミングで、無効な書き込みデータを含む部分的に無効なシリンドから、有効なデータを収集して他の領域に書き込むことにより、そのシリンドを空き領域のみからなるシリンドにするようになっている。

【0014】

【発明が解決しようとする課題】この方法によれば、パリティデータを構成するデータは、物理的トラックの長さを有するため、パリティグループを形成するまで複数のデータを保持するキャッシュの容量は大きくなる。

【0015】本発明の目的は、更新用のデータを保持するより少ない容量のキャッシュメモリを使用して、動的にアドレス変換により更新データの書き込みを行いうるデータ書き込み方法およびそのための装置を提供することを目的とする。

【0016】

【課題を解決するための手段】本発明では、複数のディスクドライブと、一つまたは複数の上位装置から転送された、それぞれ該複数のディスクドライブに新に書き込まれるべき所定の長さのデータあるいは該複数のディスクドライブに書き込み済みのデータを更新すべき該所定の長さを有するデータのいずれかからなる複数の書き込みデータを一時的に保持するキャッシュメモリとを有するディスクアレイシステムにおいて、(a) 該キャッシュメモリに保持された複数の書き込みデータから、誤り訂正用データグループを構成するためのデータとして、それぞれ該所定長を有する該所定数に等しい数の書き込みデータを取り出し、取り出された該所定数の書き込みデータから誤り訂正用のデータを生成し、(b) 該所定数の書き込みデータと該誤り訂正用のデータとを、一つの誤り訂正データグループとして、該複数のドライブの内に互いに異なるものに属する複数の空き領域に並列に書き

10

20

40

12

込み、(c) 該所定数の書き込みデータが、該複数のディスクドライブに書き込み済みのデータに対する更新データを含む場合には、その書き込み済みのデータを無効にする。

【0017】その後、適当なタイミングで、部分的に無効となったパリティグループ内の有効なデータを集めて、新たなパリティグループを生成し、空き領域に書き込み、元の部分的に無効なパリティグループ内の有効なデータをすべて無効とする。

【0018】

【作用】書き替え前のデータ、パリティデータを読みださずにパリティデータを作成でき、書き替えのための旧データ、旧パリティデータの読み出しが不要となった。

【0019】また、パリティグループを構成するデータの長さを上位装置から送られてくる一定のデータ長にしたので、パリティグループを生成するまでに書き込みデータを保持するためのキャッシュの容量が小さくてよい。

【0020】

【実施例】

(実施例1)

(1) 概要

以下本発明の一実施例を図1により説明する。

【0021】本実施例はCPU1、アレイディスクコントローラ(以下ADC)2、アレイディスクユニット(以下ADU)3により構成される。ADU3は複数の論理グループ10により構成され、個々の論理グループ10はm台のSCSIドライブ12と、各々のSCSIドライブ12とADC2を接続するドライブバス9-1から4により構成される。なお、このSCSIドライブ12の数は本発明の効果を得るには、特に制限は無い。この論理グループ10は障害回復単位で、この論理グループ10内のSCSIドライブ12は、m-1個のデータとそれらから生成したパリティデータとからなる誤り訂正データグループ(以下では、簡単化のためにパリティグループと呼ぶ)を保持する。

【0022】次にADC2の内部構造について図1を用いて説明する。ADC2はチャネルバスディレクタ5と2個のクラスタ13とバッテリバックアップ等により不揮発化された半導体メモリであるキャッシュメモリ7により構成される。このキャッシュメモリ7にはデータとアドレス変換用テーブルが格納されている。このキャッシュメモリ7およびその中のアドレス変換用テーブルはADC2内の全てのクラスタにおいて共有で使用される。クラスタ13はADC2内において独立に動作可能なバスの集合で、各クラスタ13においては電源、回路は全く独立となっている。クラスタ13はチャネル、キャッシュメモリ7間のバスである、チャネルバス6と、キャッシュメモリ7、SCSIドライブ12間のバスであるドライブバス6-1から4が、それぞれ、2個

50

ずつで構成されている。それぞれのチャネルバス6-1から4とドライブバス8はキャッシュメモリ7を介して接続されている。CPU1より発行されたコマンドは外部インターフェースバス4を通ってADC2のチャネルバスディレクタ5に発行される。ADC2は2個のクラスタ13により構成され、それぞれのクラスタは2個のバスで構成されるため、ADC2は合計4個のバスにより構成される。このことから、ADC2ではCPU1からのコマンドを同時に4個まで受け付けることが可能である。そこで、CPU1からコマンドが発行された場合ADC2内のチャネルバスディレクタ5によりコマンドの受付が可能かどうか判断する。図2は図1のチャネルバスディレクタ5と1クラスタ13-1内の内部構造を示した図である。図2に示すように、CPU1からADC2に送られてきたコマンドはインターフェースアダプタ15により取り込まれ、マイクロプロセッサ(MP)20はクラスタ内の外部インターフェースバス4の中で使用可能なバスがあるか調べ、使用可能な外部インターフェースバス4がある場合はMP20はチャネルバススイッチ16を切り換えてコマンドの受け付け処理を行ない、受け付けられない場合は受付不可の応答をCPU1へ送る。

【0023】レベル3のRAIDでは、基本的には、複数の書き込みデータから誤り訂正用のデータ、例えばパリティデータを生成し、上記複数の書き込みデータと誤り訂正用のデータとを、一つの誤り訂正データグループとして複数のドライブに分散して記憶される。

【0024】誤り訂正用のデータには、パリティデータ以外のデータも使用可能である。しかし以下では、表現の簡略化のために、パリティデータ以外のデータを使用する場合を含めて、誤り訂正用のデータを単にパリティデータと呼び、誤り訂正データグループをパリティグループと呼ぶ。

【0025】本実施例の概略的な動作は以下のとおりである。

【0026】CPUから転送された書き込みデータを上記第1の領域に書き込む方法は、基本的には、従来のレベル5のRAIDによる方法に従う。

【0027】すなわち、複数の書き込みデータをキャッシュメモリ7に保持し、保持された所定数のそれぞれ物理ドライブのレコード長に等しい長さの書き込みデータからパリティデータを生成し、これらのデータをパリティグループとして、いずれかの論理グループ12内の複数のドライブに設けられた第1の領域内に分散して書き込む。

【0028】書き込み済みのデータの読み出しをCPUから要求されたときには、そのデータのみを選択的に読み出し、キャッシュメモリ7を介してCPUに送る。

【0029】書き込み済みのデータを更新するデータがCPUから転送されたときには、その書き込み済みのデ

ータとその書き込み済みのデータが属するパリティグループ内のパリティデータをキャッシュメモリ7に読み出し、これらの読み出されたデータと更新用のデータから、新たなパリティデータを生成し、この更新用のデータと新たなパリティデータでもって、古い書き込みデータと古いパリティデータを書き換える。従って、元のパリティグループ内のデータが部分的に更新されることになる。

【0030】これに対して、CPUから転送された書き込みデータを上記第2の領域に書き込む方法は、本実施例で特徴的である。すなわち、CPUから転送され、キャッシュメモリ7に保持された書き込みデータは、いずれかの論理グループに書き込まれているデータを書き換える更新用のデータであるか否かにかかわらず、それぞれレコード長に等しい、該所定数分集められ、それらからパリティデータが生成され、新たなパリティデータグループとして、いずれかの論理グループ12内の複数のドライブ内の設けた第2の領域内の空き領域に分散して記憶される。この結果、従来技術の問題のところで述べた、書き込みデータを更新するときのオーバヘッドは減少する。

【0031】この複数の書き込まれたデータのうちのいずれか一つが、書き込み済みのデータを更新するデータであるときには、その書き込み済みのデータが無効にされる。この結果、その書き込み済みのデータが属していたパリティグループは、部分的に無効にされることになる。

【0032】以上の書き込み処理を複数のパリティグループに対して行なうと、論理グループ内の空き領域が減少する。このため、本実施例では、部分的に無効となつた、パリティグループ内の有効なデータをキャッシュメモリ7に集め、それらから新たなパリティグループを作り、空き領域に記憶する。この収集が完了した、部分的に無効なパリティグループは、空き領域として新たなパリティグループの書き込みにその後使用される。

【0033】このように、第2の領域にCPUから指定されたあるアドレスを有するデータを書き込んだ後、そのアドレスを有する別の書き込みデータがCPUから転送された場合、後のデータは、前のデータが書き込まれた、ドライブ内位置とは異なる位置に記憶される。つまり、CPUが指定するアドレスを動的にドライブ内アドレスに変換する。従って、以下では、第2の領域を、動的アドレス変換をする領域とも呼ぶ。

【0034】また、本実施例では、更新されるデータはドライブからは読み出されないで、無効にされる。このため、前述の国際公開公報に記載された動的変換とはことなり、データの更新を高速に行ない得る。

【0035】(2) アドレス変換テーブル

本実施例ではADU3を構成するSCSIドライブ12はSCSIインターフェースのドライブを使用する。C

PU1をIBMシステム9000シリーズのような大型汎用計算機とした場合、CPU1からはIBMオペレーティングシステム(OS)で動作可能なチャネルインターフェースのコマンド体系にのっとってコマンドが発行される。そこで、SCSIドライブ12をSCSIドライブを使用した場合、CPU1からのコマンドを、SCSIインターフェースのコマンド体系にのっとったコマンドに変換する必要が生じる。この変換はコマンドのプロトコル変換と、アドレス変換に大きく分けられる。

【0036】以下にアドレス変換用のテーブルについて説明する。

【0037】以下では、CPU1から転送されるデータ長はドライブの1セクタ長に等しいか、又はそのn倍(nは1より大きい整数)とする。セクタ長のn倍の書き込みデータ(レコード)を本実施例では、セクタ長のm倍(mはnより小さい整数)に等しい長さの複数のデータ(ブロック)に分割して処理する。しかし、以下では、簡単化のために、CPU1から転送されたデータのレコード長が常に一定な場合についてのみ説明する。CPU1から指定されるアドレスは、図13に示すようにデータが格納されているトラックが所属するシリンドの位置とそのシリンド内において当該データが格納されているトラックを決定するヘッドアドレスと、そのトラック内のレコードの位置を特定する。具体的には要求データが格納されている当該ドライブの番号(ドライブ番号)と当該ドライブ内のシリンド番号であるシリンドアドレス(CC)とシリンドにおいてトラックを選択するヘッドの番号であるヘッドアドレス(HH)とレコードアドレス(R)からなるCCHHRである。従来のCKDフォーマット対応の磁気ディスクサブシステム(IBM3990-3390)ではこのアドレスに従ってドライブへアクセスすれば良い。しかし、本実施例では複数のSCSIドライブ12により従来のCKDフォーマット対応の磁気ディスクサブシステムを論理的にエミュレートする。つまり、ADC2は複数のSCSIドライブ12が、従来のCKDフォーマット対応の磁気ディスクサブシステムで使用されているドライブ1台に相当するようにCPU1にみせかける。このため、CPU1から指定してきたアドレス(CCHHR)をSCSIドライブのアドレスにMP20が変換する。このアドレス変換には図3に示すようなアドレス変換用のテーブル70(以下アドレステーブルとする)が使用される。ADC2内のキャッシュメモリ7には、その内部の適当な領域にこのアドレステーブル70が格納されている。本実施例では、CPU1が指定してくるドライブはCKDフォーマット対応の単体ドライブである。しかし、本発明ではCPU1は単体と認識しているドライブが、実際は複数のSCSIドライブにより構成されるため、論理的なドライブとして定義される。このため、ADC2のMP20はCPU1より指定してきたCPU指定アドレス750

1(これはドライブ番号74とCCHHR75からなる)をSCSIドライブ12に対するSCSIドライブアドレス72(これはSCSIドライブ番号77とそのSCSIドライブ内のアドレス(以下SCSI内Addrとする)78からなる)に変換する。アドレステーブル70はCPU1が指定するCPU指定アドレス71とそれに対応する、実際にデータが格納されているSCSIドライブ12内のアドレス(SCSIドライブアドレス)72と、そのデータに対応したパリティデータの格納されているパリティドライブアドレス73とキャッシュメモリ7内のアドレス(キャッシュアドレス)81と、キャッシュメモリ7内に当該データが存在するかどうかのキャッシュフラグ82が格納される。このキャッシュフラグ82はオン(1)の場合キャッシュメモリ7内にデータが存在していることを示し、オフ(0)の場合はキャッシュメモリ7内にデータが存在していないことを示す。

【0038】また、アドレステーブル70には動的アドレス変換を行なう場合の図4に示す動的アドレス変換テーブル90へのポインタ(DMポインタ)76が格納される。SCSIドライブアドレス72はデータの格納されているSCSIドライブの番号(SCSIドライブ番号)77とSCSI内Addr78(これはSCSIドライブ内のシリンドアドレス、ヘッドアドレス、レコード番号、セクタ数からなる)により構成されており、パリティドライブアドレス73にはそのデータに対応したパリティデータの格納されているSCSIドライブの番号(パリティドライブ番号)79とパリティデータの格納されているSCSIドライブ内のアドレス(以下パリティデータ内Addrとする)80(これはパリティドライブ内のシリンドアドレス、ヘッドアドレス、レコード番号、セクタ数からなる)が格納される。なお、アクセスフラグ84は、実施例2で使用するものである。

【0039】キャッシュアドレス81はDMポインタ76により対応するSCSIドライブアドレス72により決定したデータが、キャッシュメモリ7内に有る場合は、このデータのキャッシュメモリ7内のアドレスを登録しておく。また、この様にキャッシュメモリ7内に当該データが格納されている場合は、アドレステーブル70と同様にキャッシュフラグ82をオン(1)とする。無効フラグはDMポインタ76に対応するSCSIドライブアドレス72により決定したデータが有効であるか無効であるかを示すフラグで、無効である場合はオン(1)が登録される。ドライブフラグはドライブ内に書き込まれているかどうかを示すフラグで、オン(1)の場合はこのデータはドライブに書き込まれており、オフ(0)の場合は、まだドライブには書き込まれていないことを示す。

【0040】なお、アドレステーブル70と動的アドレス変換テーブル90はシステムの電源をオンした時に、

MP 2 0により論理グループ1 0内のある特定のSCSIドライブ1 2から、キャッシュメモリ7にCPU1の関与なくして自動的に読み込まれる。一方、電源をオフする時は、MP 2 0によりキャッシュメモリ7内のアドレステーブル7 0を、読み込んできたSCSIドライブ1 2内の所定の場所にCPU1の関与なくして自動的に格納する。

【0041】(3)動的アドレス変換を行う領域とその他の領域の確保

以下、本実施例の処理を詳細に説明する。

【0042】まず、ユーザは、ディスクアレイを使用する前に動的アドレス変換を行なうデータを格納する領域と、通常のレベル5で処理する領域をCPU1を介して設定する。

【0043】本実施例による動的にアドレス変換を行なうと、空き領域を予め用意しておかなければならず、また、処理中に空き領域が無くなかった場合は詰め替えを行い、空き領域を確保しなければならない。このため、本実施例では非常にランダム性の高いアクセスが生じるデータについては詰め替えが頻繁に発生するため、なるべく動的にアドレス変換を行なわない。そこで、ユーザは動的にアドレス変換を行うことで性能向上が図れるシーケンシャルデータの量を予め調査し、動的アドレス変換を行なう領域を設定する。設定方法は、初期設定時にユーザは確保したい領域をCPU指定アドレスの特定の範囲の形でADC2に申請する。ADC2ではMP20が図3、4に示すように、このユーザが要求する動的アドレス変換を行なう領域の大きさに見合う領域を、SCSIドライブ1 2内に確保した領域のアドレス(SCSI内Add7 8)を動的アドレス変換テーブル9 0に登録する。MP1 2 0はアドレステーブル7 0の一つの行と動的アドレス変換テーブル9 0の一つの行との間にリンクをはるために、それぞれに同じ値のDMピント7 6をセットする。上述のCPU指定範囲に属するCPU指定アドレスを特定アドレスと呼ぶ。後にCPU1からデータの書き込みを要求するときに、動的アドレス変換を行う場合、前述のCPU指定範囲に属する特定アドレスを書き込み先として指定したとき、このCPU指定アドレスは、アドレステーブル7 0内のDMピント7 6が設定されている一つの行に登録される。上記CPU指定範囲に属しないCPU指定アドレスが書き込み要求で指定されたときには、アドレステーブル7 0内の、DMピント7 6が設定されていない一つの行にこのCPU指定アドレスが登録される。また、動的アドレス変換はシーケンシャルデータのみでなく、データ圧縮を行なうデータに関する動的アドレス変換を行なう領域に格納する。

【0044】動的アドレス変換を行う領域は、各論理グループ1 0を構成する複数のドライブ内の同じアドレス範囲の領域に分散して構成する。それぞれの領域は、そ

れぞれセクタ単位の大きさを有する複数の領域から構成してもよく、あるいはシリンドラ単位の大きさを有する複数の領域から構成してもよい。この様に領域を設定された後、実際の処理としては、CPU1が指定してきたアドレス(CPU指定アドレス7 1)をMP20がキャッシュメモリ7内のアドレステーブル7 0によりアドレス変換する際、CPU指定アドレス7 1が動的アドレス変換を行なう領域のアドレス(特定アドレス)かどうかをDMピント7 6を参照し、特定アドレスのデータであれば、動的アドレス変換を行なうように処理し、特定アドレスでないアドレスのデータであれば通常のレベル5のように処理する。

【0045】(4)データ書き込み動作の概要

以上のような動的アドレス変換を行なう領域とレベル5の領域を設定した後、以下のようなデータの書き込み処理を行なう。

【0046】CPU1から書き込み命令が発行されたとする。まず、ADC2のいずれかのMP20はCPU1からCPU指定のドライブ番号と、書き込みデータとCPU指定アドレス7 1を指定する書き込みコマンドを受け取った後、そのMP20が所属するクラスタ1 3内の各チャネルバス6において処理可能かどうか調べ、可能な場合は処理可能だという応答をCPU1へ返す。CPU1では処理可能だという応答を受け取った後にADC2へデータを転送する。この時、ADC2ではMP20の指示によりチャネルバスディレクタ5において、チャネルバススイッチ1 6が当該外部インターフェースバス4とインターフェースアダプタ1 5を当該チャネルバス6と接続しCPU1とADC2間の接続を確立する。CPU1とADC2間の接続を確立後CPU1からのデータ転送を受け付ける。CPU1から転送されてきたデータはMP20の指示により、チャネルインタフェース2 1によりプロトコル変換を行ない、外部インターフェースバス4での転送速度からADC2内の処理速度に速度調整する。チャネルインタフェース2 1におけるプロトコル変換および速度制御の完了後、データはDCC2によるデータ転送制御を受け、圧縮回路2 7に転送される。

【0047】データ圧縮を行なうデータについては圧縮回路2 7に転送されてきたデータはMP20の指示によりデータ圧縮される。

【0048】圧縮回路2 7により圧縮されたデータは、チャネルアダプタ2 4に転送され、そのアダプタ2 4によりキャッシュメモリ7内に格納される。データ圧縮を行なわないデータはDCCから圧縮回路2 7内のスルーパスを通りチャネルアダプタ2 4に転送され、圧縮データと同様にキャッシュメモリ7内に格納される。書き込みデータを格納するキャッシュメモリ7はバッテリ等により不揮発化されている方が好ましい。この様にキャッシュメモリ7にデータを格納したのをMP20が確認し

たら、MP 20 は書き込み処理の完了報告を CPU 1 に対し報告する。

【0049】このような動作を繰り返して、複数の書き込みデータが、キャッシュメモリ 7 に書き込まれる。この際、その書き込みデータに対して指定されたアドレスが前述の CPU 指定アドレスの特定の範囲に属するものか否かをアドレステーブル 70 に基づいて判断され、その結果を用いて、これらの書き込みデータは、動的書き込み領域に書き込むべきデータのグループとそうでないデータのグループに分かれ、キャッシュメモリ 7 内に記憶される。これらの書き込みデータは、それぞれのグループ毎に処理される。

【0050】動的アドレス変換をしない領域に書き込まれるべきデータについては以下のように処理される。

【0051】まず、それぞれの書き込みデータに対して CPU が指定したアドレスが、すでにアドレス変換テーブルに登録されているか否か、すなわち、それぞれの書き込みデータが、すでに書き込み済みのデータを更新するための書き込みデータであるかあるいはそのアドレスのデータが初めて書き込まれるものかが判断される。

【0052】書き込み済みのデータを更新しないデータは、それらが所定数に達すると、MP 20 の指示により、パリティデータ発生回路 (PG) 36 が、それらのデータからパリティデータグループを生成し、MP 20 は、それらの書き込みデータおよびパリティデータを、書き込むべき空き領域を決定する。

【0053】すなわち、いずれかの論理グループ 10 内の所定数のドライブ内の、上記領域に属する、互いにドライブ内アドレスが等しい空き領域をアドレステーブル 70 を参照して選択する。その後、アドレステーブル 70 にこれらの書き込みデータの CPU 指定アドレスを登録し、これらの書き込みデータに割り当てられた記録領域のドライブ番号 77 及びドライブ内アドレス 78、およびパリティデータに関する同様のアドレス 79、80 を登録し、キャッシュアドレス 81 を登録した上で、キャッシュフラグ 82 をオン (1) にセットする。

【0054】かかる後、これらのパリティグループを、上で割り当てられた空き領域に記憶する。

【0055】動的アドレス変換をしない領域に書き込まれるべきデータが、書き込み済みのデータを更新するデータであると判断された場合には、以下に詳しく説明するように、その書き込み済みデータとその書き込み済みのデータが属する元のパリティグループ内の元のパリティデータが読み出され、それらのデータと更新用の書き込みデータから新たなパリティデータが生成され、その更新用のデータとその新たなパリティデータでもって、元の書き込み済みのデータおよび元のパリティデータを書き換える。従って、元のパリティグループが書き換えられるのであって、新たなパリティグループは生成されない。

【0056】以下この場合の書き込み動作を図 11 を参考して説明する。

【0057】

(5) 非動的アドレス変換をする領域内のデータの更新
まず MP 1 20 がドライブインターフェース 28 に対し書き込み済みのデータとパリティデータを読みだすよう指示を出す。具体的には図 3 において CPU 指定アドレス 71 が Drive # 2, ADR 3 の位置にあるデータを書きかえる場合、MP 20 はアドレステーブル 70 において CPU 指定アドレス 71 (Drive # 2, ADR 3) に対応する各項目を調べる。図 3 において MP 20 はこの CPU 指定アドレスはすでに登録されているが DM ポインタ 76 がそのアドレスに対して登録されていないため、このデータはレベル 5 で処理するデータと判断し、SCSI ドライブアドレス 72 とパリティドライブアドレス 73 から書き込み先の SCSI ドライブ番号 77 と SCSI 内 Addr 78 とパリティドライブ番号 79 とパリティデータ内 Addr 80 を認識する。MP 20 はこの変換後のアドレスによりドライブインターフェース 28 に対し書き込み先のアドレスにすでに格納されている書き込み前のデータおよびパリティデータの格納されている SCSI ドライブ 12 に対し、当該データおよび当該パリティデータの読み出し要求を発行するよう指示を出す。ドライブインターフェース 28 からこの読み出し要求を受けた 2 つの SCSI ドライブ 12 においては当該 SCSI 内 Addr 78、およびパリティデータ内 Addr 80 に対しアクセスし、書き込み済みのデータおよびパリティデータをそれぞれの当該 SCSI ドライブからキャッシュメモリ 7 に転送する。パリティデータ生成回路 36 はこれらの書き込み済みのデータとパリティデータと、新たな書き込みデータとから MP 20 の指示で、更新後の新しいパリティデータを作成する。新たな書き込みデータと新たなパリティデータをそれぞれ更新前の書き込みデータおよびパリティデータが格納されていたアドレスに書き込む。

【0058】(6) 動的にアドレス変換をする領域内のデータの書き込みと更新

このように、動的アドレス変換をしない領域に書き込まれるべきデータは、そのデータがすでに書き込み済みのデータを更新するデータであるか否かにより異なる処理を受ける。

【0059】一方、動的アドレス変換をする領域に書き込まれるべきデータについては、以下のように処理される。

【0060】そのようなデータは、書き込み済みのデータを更新するデータであるか否かに依らずに、所定数ずつまとめて、新たなパリティグループとして論理グループ内の空き領域に記憶される。

【0061】すなわち、そのような書き込みデータが所定数に達した場合には、MP 20 の指示により、パリテ

21

イデータ生成回路36がそれらのデータからパリティデータを生成し、それらの書き込みデータと生成されたパリティデータとから、新たなパリティグループを生成する。MP20は、それらのパリティグループを書き込むべき空き領域を決定する。先の、動的アドレス変換をしない場合とことなり、空き領域は、動的アドレス変換用の領域に属する領域から選択する。

【0062】すなわち、いずれかの論理グループ10内の所定数のドライブの互いにドライブ内アドレスが等しい空き領域を動的アドレステーブル90を参照して選択する。その後、アドレステーブル70にこれらの書き込みデータのCPU指定アドレスを登録し、さらに、動的アドレステーブル90に、これらの書き込みデータに割り当てられた記録領域のドライブ番号77及びドライブ内アドレス78、およびパリティデータに関する同様のアドレス79、80を登録し、キャッシュアドレス81を登録した上で、キャッシュフラグ82をオン(1)にセットし、かつ、無効フラグ91、ドライブフラグ83は0のままにする。

【0063】しかる後、これらのパリティグループを、上で割り当てられた空き領域に記憶する。

【0064】この新たなパリティグループを構成する書き込みデータのいずれもが、書き込み済みデータを書き換えるデータでないときには、書き込み動作は以上で終了するが、それらの書き込みデータのいずれかが、すでに書き込み済みのデータを更新する場合、その書き込み済みのデータを無効にする。従って、元のパリティグループは、部分的に無効なデータを含むことになる。

【0065】以下、動的アドレス変換をする領域へのデータの書き込みの詳細を図6を参照して説明する。

【0066】例えばSCSIドライブSD#1の内に書き込み済みデータ#1、次にSD#2書き込み済みデータ#2、次にSD#3のデータ#3、次にSD#4に書き込み済みのデータ#4に対して順に書き込み命令がCPUより発行された場合を例にして説明する。

【0067】MP20はこれらのデータをキャッシュメモリ7に書き込む。このとき、キャッシュ内のアドレステーブル70、90にそれぞれのデータについての情報を登録する。なお、更新前の書き込みデータ#1～#4がキャッシュメモリ7内に保持されている場合には、キャッシュメモリ7内のそれらの旧データを書き替えた後、以下のことを行う。

【0068】論理グループ10を5台のSCSIドライブ12で構成すると仮定する。MP20は、キャッシュメモリ7に4個の書き込みデータが揃うとパリティデータの作成を行う。

【0069】すなわち、MP20は、動的アドレス変換テーブル90において、無効フラグ91がオフ(0)でキャッシュフラグ82がオン(1)でドライブフラグ83がオフ(0)のデータを探す。MP20はこの条件を

10

22

満足するデータの数を調べその数が4に達し次第、その4つの書き込みデータについて、パリティデータ生成回路(PG)36にパリティデータの作成を指示する。MP20は、これら4つのデータと新たに生成したパリティデータとを、一つの新たなパリティグループとして論理ドライブに並列に書き込む。このために、MP20はまず書き込み先のアドレスを探す。つまり、各SCSIドライブ12においてデータ#1から4とパリティデータが全て書き込み可能なスペースをさがす。具体的には、MP20は動的アドレス変換テーブル90において、論理グループを構成するSCSIドライブ12に対し、同一SCSI内Addr78の無効フラグ91がオンになっている領域を探す。この場合、対応するパリティデータが有効か否かは問わない。

【0070】以下ではデータ#1から4のCPU指定アドレス71が、図4に示す動的アドレス変換テーブル90においてデータ#1がDM a-2、データ#2がDM b-3、データ#3がDM c-1、データ#4がDM d-3に対応するとする。

【0071】図4の場合SCSIドライブ#1から5が論理グループを構成するとしてドライブSD#1から4のアドレスSADR4の領域は全て無効フラグ91がオンになっており、データ#1から4を書き込める空き領域と判断する。

【0072】そこで、ユーザがデータ#1の書き込み先に指定してきたCPU指定アドレス71(特定アドレス)に対し、MP20はDMポインタ76により動的アドレス変換テーブル90のDM a-2の項目で、ドライブSD#1のSADR2に書き込む様に変換されるが、動的アドレス変換を行なうため、書き込みスペースの判断後、新たな書き込み先と決定した、SCSIドライブSD#1内のSADR4に書き込むことを決定する。同様に、MP20はドライブSD#2のアドレスSADR3のデータ#2、SD#3のSCSIドライブ12のSADR1のデータ#3、ドライブSD#4のSADR3のデータ#4をそれぞれ更新する書き込みデータ#2、#3、#4も、SCSIドライブSD#2～SD#4のアドレスSADR4に書き込むことを決定する。

【0073】この様に、MP20が書き込み先のアドレスを決定した後、MP20が上記書き込み可能なスペースのSCSI内Addr78(SADR4)に対し、ドライブインターフェース28に、各SCSIドライブSD#1～SD#5への書き込み要求を発行するように指示する。ドライブインターフェース28ではSCSIの書き込み処理手順に従って、当該SCSIドライブに対し書き込みコマンドとSCSI内Addr78(SADR4)をドライブユニットバス9を介して発行する。ドライブインターフェース28から書き込みコマンドを発行された各SCSIドライブ12においては指示SCSI内Addr78(SADR4)へシーク、回転待ちのアクセス処理を

50

行なう。当該SCSIドライブ12におけるアクセス処理が完了した後、キャッシングアダプタ14はキャッシングメモリ7からデータを読み出してドライブインターフェース28へ転送する。ドライブインターフェース28は転送されてきたデータをドライブユニットバス9を介して該当するSCSIドライブ12へ転送する。データの該当SCSIドライブ12のSCSI内Addr78(SADR4)への書き込みが完了すると、当該SCSIドライブ12はドライブインターフェース28に完了報告を行ない、ドライブインターフェース28がこの完了報告を受け取ったことを、MP20に報告する。この時、MP20は、この書き込みデータをキャッシングメモリ7上に残さない場合は、この報告を元にアドレステーブル70のキャッシングフラグ82をオフにする。パリティデータに関してもデータと同様にパリティドライブアドレス73に従って書き込まれる。また、この様に書き込みアドレスを動的に変換した場合は、SCSIドライブ12からの完了報告をドライブインターフェース28が受け取った際に、アドレステーブル70のユーザが書き込み先として指定したCPU指定アドレス71に対するDMポインタ76の値を、動的アドレス変換テーブル90における書き込み後のSCSIドライブアドレス72に対応するDMポインタ76の値に変更する。この時、同時に動的アドレス変換テーブル90の各項目も変更する。具体的には図3のアドレステーブル70において、CPU1がデータ#1をDrive#1のADR1に書き込むように指示した場合、このデータ#1を上記のように動的アドレス変換した場合、書き込み前はDMポインタ76はDMA-2となっているため、動的アドレス変換テーブル90において、SCSIドライブSD#1のアドレスSADR2が対応していた。動的アドレス変換後により、データ#1の書き込み後は、MP1-20はアドレステーブル70のDrive#1, ADR1のDMポインタ76をDMA-2からDMA-4に変更する。これと同時にMP20は動的アドレス変換テーブル90においてDMA-a-4の無効フラグ91をオフ(0)、ドライブフラグ83をオン(1)にし、書き込みデータをキャッシング7内に残すときは、キャッシングアドレス81にアドレスを登録しキャッシングフラグ82をオン(1)のまととする。将来、ユーザがこのデータに読み出し要求を発行した場合は前のアドレスにはアクセスせず新しく格納されたアドレスにアクセスすることになる。

【0074】また、同時に動的アドレス変換テーブル90において、書き込み前のDMA-a-2の無効フラグ91をオン(1)とし、キャッシングフラグ82、ドライブフラグ83をオフ(0)とする。この領域は後の別な書き込みのスペースとして使用される。

【0075】以上のようなドライブへの書き込み動作は、論理グループ10を構成する各SCSIドライブ12において並列に行なう。以下では以上の書き込み動作

を一括書き込みとも呼ぶ。

【0076】この様に動的アドレス変換をする領域に対しては、書き込みデータが書き込み済みのデータを更新するか否かに関係なく、書き込みデータをキャッシングメモリ7に溜め、溜められた所定数の書き込みデータによりパリティデータを作成し、書き込みデータとパリティデータを複数のSCSIドライブ12へ並列に書き込む。従来のレベル5のようなユーザが指定した書き込み先のアドレスにすでに格納されている、書き込み前のデータとパリティデータに対する読みだしは行なわない。このため書き込み時のオーバヘッドを削減することが可能となる。

【0077】しかし、書き込みデータについて動的にアドレスを変更し格納していくと、無効フラグ91がオン(1)になっている領域が分散されるため、データ格納効率(SCSIドライブに格納可能なデータ容量に対する、実際に格納するデータ容量の割合)が低下し、さらにこの様な状態が進行すると、動的アドレス変換による一括書き込みを行なう空き領域が無くなってしまう。これに対しては、後述するデータのつめ替えで対処する。

【0078】なお、本実施例では同じドライブ内アドレスを有する複数の領域にデータを並列に格納するため、論理グループ10内の各SCSIドライブ12の回転を全て同期させる方が望ましい。

【0079】(7)障害回復

いくつかの無効データが存在する場合において、論理グループ10内のSCSIドライブに障害が発生した場合、障害SCSIドライブを除いた残りのSCSIドライブ12内のデータおよびパリティデータにより障害SCSIドライブ内のデータを復元することが可能となる。このような障害回復を行なう場合、各SCSIドライブ12内のキャッシングメモリ7内の対応する無効フラグ91がオンになっているデータも有効なデータとして読みだして障害データの復元に使用する。無効フラグ91がオンになったデータは更新前の古いデータであるが、それが属していたパリティグループが生成されたときの値を有している。したがって、障害ドライブのデータの回復には使用できる。このため、本実施例では、動的アドレス変換領域に書き込み済みのデータが更新された場合、そのデータに対して無効フラグをつけるが、データは消去しないのは、そのためである。

【0080】

(8)動的アドレス変換領域と非動的アドレス変換領域まず、データ格納領域について説明する。本実施例ではSCSIドライブ12内の動的アドレス変換を行なう領域に実際に格納するデータの量は、動的アドレス変換を行なう領域に格納可能な最大容量の1/Vとする。これは、書き込みデータのアドレスを動的に変換して一括書き込みを行なうため、一括書き込みを行なう領域(書き替え領域)が必要となるからである。つまり、SCSI

ドライブ12の動的アドレス変換を行なう領域に書き込み可能な最大容量のデータを格納してしまうと、後に書き込み要求が発行された場合、一括書き込みを行なおうとしても書き込む場所が無くなってしまうためである。大型汎用計算機では、ユーザは予め自分の使用可能な領域(容量)を確保しておく。ユーザがデータの読みだし、書き込みを行なう場合は、この容量に見合う領域内で処理しなければならない。もし、この使用可能な領域(容量)を越えた場合は新たに領域(容量)を確保しなければならない。そこで、先に示したようにユーザがCPU1により動的アドレス変換を行なう領域を設定する際、予め後の書き込み要求のために書き替え領域を見込んで、領域を確保しておく((V-1)/V)。

【0081】この予め確保しておく書き替え領域の割合は、動的アドレス変換をシーケンシャルアクセスされるデータのみで行なうのではなく、一部ランダムアクセスされるデータについても行なう場合、ランダムデータに関し、ランダムなアドレスにアクセスする読み出し/書き込み要求が多い場合は1/Vの値を小さくし、比較的同じようなアドレスにアクセスする読み出し/書き込み要求が多い場合は1/Vの値を大きくするようにVの値を変えることが望ましい。さらに、前者の場合、キャッシュメモリ7から早くSCSIドライブにデータを移し、後者の場合は逆にデータがキャッシュメモリ7内に長く滞留させることが望ましい。1/Vを変更する方法には2つの方法がある。一つは、ユーザが予め格納するデータについて、ランダムなアクセスの量を認識し、格納するデータの割合(1/V)を決める。ユーザは自分のデータの特性から判断し、自分が確保した領域(容量)の中で、実際に格納するデータの割合を決めるのである。例えば、同じデータに対して次々と書き換えを行なうアクセスの割合が、全アクセスに対して約30%程度なら、1/V=1/2程度に設定し、予めユーザが確保した領域(容量)の1/2のみユーザはデータを格納しておく。もう一つは、ユーザが予め領域(容量)を確保するのではなく、ADC2内のMP20が一定時間アクセスされたアドレスの履歴を記録しておき、そのアドレス履歴データからランダムなアクセスの割合を求め、ADC2のMP20が格納するデータの割合(1/V)を自動的に割り当てる。例えばSCSIドライブ12に格納可能なデータ量の4/5のデータが格納されており、このデータに対してランダムなアクセスの割合が増加した場合、別の論理グループに随時データを移し、MP120は1/Vを小さくする。

(9) 部分無効パリティグループの詰め替え

しかし、このように予め書き替え領域を確保しておいても、SCSIドライブ12内に格納できるデータの容量は有限であるため、一括書き込みを繰り返した場合、書き替え領域が無くなる。そこで、SCSIドライブ12内の有効なデータをまとめ、新たに書き替え領域を作

成する必要が生じる。この方法について次に説明する。【0082】図7は各SCSIドライブ12内の同一SCSI内Addr78のデータを抜き出したものである。論理グループ10内における各SCSIドライブ12の同一SCSI内Addr78(同一行)のデータつまり、同一のパリティグループを形成すべきデータにより、パリティデータが作成される。例えばSADR aのアドレスを考えると、SCSIドライブ#1のデータ1、SCSIドライブ#2のデータ#5、SCSIドライブ#3のデータ#9、SCSIドライブ#4のデータ#13によりパリティデータ#1が作成される。図7ではSADR aのデータグループにおいて無効フラグ91がオフのデータはデータ#9のみである。前述したように無効とされたデータ#1、#5、#13は障害発生時の障害ディスク内のデータ復元用データとして残されている。SADR bについては無効フラグ91がオフのデータはデータ#10、#14の2個で、SADR cではデータ#3、#7、#15の3個、SADR dはデータ#4、#8、#12、#16と全て無効フラグ91がオフとなっている。MP20はキャッシュメモリ7内のアドレステーブル70および動的アドレス変換テーブル90を常に監視しており、動的アドレス変換テーブル90内のSCSIドライブアドレス72と無効フラグ91を見て、同一パリティグループに属する、すなわち、同一SCSI内Addr78に対応した、無効フラグ91のオン(1)となっている領域の数を認識しこの領域の数が予め設定しておいた数より多くなると、詰め替えを行う。例えば、同一SCSI内Addr78に対し無効フラグ91がオン(1)になった領域の数が3個より多くなると詰め替えを行うように予め設定しておいた場合、図7のSADR aを図4のSADR 3とすると、このように同一SCSI内Addr78に対し無効フラグ91が3個オン(1)になると、MP20は詰め替えを行う。また、この詰め替え処理はユーザの指示で行なうことも可能とする。

【0083】具体的には、MP20は、その部分無効パリティグループ内の有効なデータ、つまり、今の例では、図4のドライブSD#3のアドレスSADR 3のデータに対し、擬似的な読みだし要求を発行し、このデータをキャッシュメモリ7に擬似的に読みだしキャッシュメモリ7内に溜め、CPU1から転送された他の書き込みデータまたは同様に他の行から擬似的に読みだされたデータととからなる4つのデータを組にし、それらから新たにパリティデータを作成し、動的アドレス変換における書き込みと同時に書き込み用空きスペースを探し、それらの領域に並列に格納する。なお、データ#9がキャッシュメモリ7に読み込まれた段階で、MP20はキャッシュメモリ7内のアドレステーブル70内のキャッシュアドレス81を変更しキャッシュフラグ82をオンにする。また、書き替え領域への格納後は、動的アドレ

27

ス変換の書き込みと同様にアドレステーブル70および動的アドレス変換テーブル90の各項目を変更する。

【0084】また、別の詰め替え方法としては、上述したように部分無効パリティグループが詰め替え条件を満すとすぐ行うというのでなく、そのようなデータに読み出し要求がCPU1よりその後発行され、SCSIドライブ12から読み出された時に行う方法がある。すなわち、上位装置にそのデータを転送すると同時にキャッシュメモリ7内に溜め、書き込みデータまたは同様に他の行から吸い上げられたデータと新たにパリティデータを作成し、これらのデータパリティデータをパリティデータを格納するSCSIドライブ12のそれぞれの書き替え領域に格納する。

【0085】詰め替えを起動する契機としてはさらに次のものでもよい。

【0086】すなわち、部分的に無効となったパリティグループの数が、予め決めた限界数を越えたときに行なう。この際、無効となったデータの数が一つでも含んでいるようなパリティグループの数をカウントして、そのカウント値が所定数を越えたときに、詰め替えを行なう。この方法では、計数操作が簡単である。

【0087】これに対して、所定数以上の無効となったデータを含んでいるパリティグループの数をカウントして、そのカウント値が所定数を越えたときに、詰め替えを起動する方法でもよい。この方法では、詰め替えを行なう回数を制限できる。

【0088】さらに別の方として、空き領域の容量が所定値以下になったときに行なう方法でもよい。

【0089】さらに、より望ましくは、CPU1を介してユーザにより確保されたデータ書き込み容量のうち、空き領域の容量の占める割り合いが、所定値以下になったときに行なう方法でもよい。

【0090】なお、書き替え領域に格納する段階で書き込み時と同様にアドレステーブル70および動的アドレス変換テーブル90を更新する。このようにして行内の全てのデータの無効フラグ91がオンになると、その行を空き領域（書き替え領域）とする。

【0091】以上の詰め替えは、以上の二つの方法のいずれを探るにしても論理グループ10に発行される読み出し／書き込み要求が比較的少ない時、特に書き込みが少ない時に使うとシステムの読み出し／書き込み要求処理効率の低下が少なく効果的である。

【0092】(10) データの読み出し

MP20が読み出し要求のコマンドを認識すると、CPU1から指定されたCPU指定アドレス71（ドライブ番号74とCCHHR75）からMP20はアドレステーブル70を参照し、そのアドレスに対してDMポインタが登録されているか否かによりそのアドレスに対し動的アドレス変換を行なっているかを判断し、行なっていない場合は、引き続きアドレステーブル70により、当該

10

20

30

40

50

28

データのSCSIドライブアドレス72へアドレス変換を行なう。同時にキャッシュメモリ7内に当該データが存在するかどうかアドレステーブル70のキャッシュフラグ82を調べ、判定する。一方、アドレステーブル70のDMポインタ76に登録されており、動的アドレス変換を行なっている場合は、動的アドレス変換テーブル90により、アドレス変換を行なう。但し、無効フラグがオフであるドライブ内アドレスを利用する。これによりCPU指定アドレスが複数個テーブル90に登録されている場合でも、最近に書き込まれたデータを読み出せるようになる。

【0093】キャッシュヒット時は、MP20はアドレステーブル70又は動的アドレス変換テーブル90によりCPU1から指定してきたCPU指定アドレス71（ドライブ番号74とCCHHR75）をキャッシュメモリ7のアドレス（キャッシュアドレス81）に変換しキャッシュメモリ7へ当該データを読み出しに行く。具体的にはMP20の指示の元でキャッシュアダプタ回路24によりキャッシュメモリ7から当該データは読み出される。キャッシュアダプタ24はキャッシュメモリ7に対するデータの読みだし、書き込みをMP20の指示で行なう回路で、キャッシュメモリ7の状態の監視、各読みだし、書き込み要求に対し排他制御を行なう回路である。キャッシュメモリ7内には圧縮回路27によりデータ圧縮されているデータも格納されているため、キャッシュアダプタ24により読み出されたデータは圧縮回路27に送られる。データ圧縮されていないデータは圧縮回路27をそのまま通過するが、データ圧縮されているデータは、圧縮回路27によりCPU1から転送されてきた圧縮前の元のデータに伸長される。圧縮回路27を通過したデータはデータ制御回路（DCC）22の制御によりチャネルインターフェース回路21に転送される。チャネルインターフェース21ではCPU1におけるチャネルインターフェースのプロトコルに変換し、チャネルインターフェースに対応する速度に速度調整する。このプロトコル変換および速度調整後は、チャネルバスディレクタ5において、チャネルバススイッチ16が外部インターフェースバス4を選択しインターフェースアダプタ15によりCPU1へデータ転送を行なう。

【0094】一方、キャッシュミス時はMP20はドライブインターフェース28に対し、SCSIドライブアドレス72に従って当該ドライブ12への読み出し要求を発行するように指示する。ドライブインターフェース28ではSCSIインターフェースの読み出し処理手順に従って、読み出しこmandoをドライブユニットバス9-1、または9-2を介して発行する。ドライブインターフェース28から読み出しこmandoを発行された当該ドライブ12においては、指示されたSCSIドライブ内のアドレス（SCSI内Add78）へシーク、回転待ちのアクセス処理を行なう。当該ドライブ12における

アクセス処理が完了した後、当該ドライブ12は当該データを読み出しドライブユニットバス9を介してドライブインターフェース28へ転送する。ドライブインターフェース28では転送されてきた当該データをドライブ側のキャッシュアダプタ回路14に転送し、キャッシュアダプタ14ではキャッシュメモリ7にデータを格納する。この時、このキャッシュアダプタ14はキャッシュメモリ7にデータを格納することをMP20に報告し、MP20はこの報告を元にアドレステーブル70または動的アドレス変換テーブル90のユーザが読みだし要求を発行したCPU指定アドレス71に対応したキャッシュフラグ82をオンにし、キャッシュメモリ7内のアドレスを以下のように更新する。CPU1から読みだし要求先のアドレスとして指定してきた、ドライブ番号とCCHHRに対し、アドレステーブル70内のCPU指定アドレス71（ドライブ番号74とCCHHR75）を探し、そのCPU指定アドレス71に対応して、当該データを格納したキャッシュメモリ7内のアドレスをキャッシュアドレス81に書き込む。これと同時にキャッシュフラグ82をオンとする。キャッシュメモリ7にデータを格納し、アドレステーブル70または動的アドレス変換テーブル90のキャッシュフラグ82をオンにし、キャッシュメモリ7内のアドレスを更新した後は、キャッシュヒット時と同様な手順でCPU1へ当該データを転送する。

【0095】(11) 変形

本実施例ではパリティデータを作成する単位は、論理グループ10を構成する各SCSIドライブ12について、同一アドレス（シリンドアドレス、ヘッドアドレス、レコードアドレスが全て等しい）とした。しかし、データ格納効率を向上させるため、同一アドレスに限定せず、インデックスからの距離が等しければ（レコードアドレスが等しい）シリンドアドレス、ヘッドアドレスは異なっていてもパリティデータを作成できる。このようにしても本実施例の動作は可能となる。

【0096】(12) 実施例1のまとめ

以上述べたように、本実施例では、動的にアドレス変換をする領域へのデータの書き込みにあっては、複数の書き込みデータからパリティデータを生成し、それらの書き込みデータと生成したパリティデータとによりパリティグループを生成し、それらを空き領域に書き込む。この書き込みは、各書き込みデータがすでに書き込み済みのデータを更新するデータである場合にも適用される。

【0097】従来のレベル5のRAIDによる書き込みのときには、書き込み済みのデータと書き込み済みの対応するパリティデータをドライブから読み出し、これらと更新用のデータとから、新たなパリティデータを生成し、この生成されたパリティデータとその更新用のデータとを旧のパリティデータが保持されている領域及び旧の書き込みデータが保持されている領域に書き込む。こ

れらの二つの領域からの旧データの読み出し及びにそれらの領域への新データの書き込みのために、ドライブの回転待等の待時間がオーバヘッドとして大きい。

【0098】本実施例では、以上に述べた書き込み方法により、以上の従来例で必要であった、ドライブからの旧データの読み出し、そこへの新データの書き込みが必要でない。そのため、以上のようなオーバヘッドが生じない。

【0099】さらに、本実施例では、ある一つのパリティグループを構成する複数のデータの一部が書き換えられたとき、そのパリティグループのいずれのデータもドライブから読み出さない。つまり、書き換えられたデータは無効とし、それ以外のデータはそのまま有効なデータとしてドライブに保持する。この結果、一部のデータの書き換え時の処理が早くなる。

【0100】さらに、本実施例では、一つのパリティグループを構成する複数のデータの長さを上位装置から送られてくるデータの一定長に等しくした。この結果、いずれかのパリティグループ内のデータが部分的に無効にされたとしても、そのパリティデータ内の全てのデータが無効となるケースが増大する。このように、完全に無効となった領域は空き領域として詰め替えをすることなくそのまま使用できる。従って、部分的に無効な状態のままでいるパリティグループは、より長いデータからパリティデータを構成した場合よりも少くなり、結果として、詰め替えを要する部分的に無効なパリティグループの数は少なくなる。

【0101】さらに、本実施例では、一つのパリティグループを構成する複数のデータの長さを上位装置から送られてくるデータの一定長に等しくした結果、上位装置から転送された複数の書き込みデータを、それらからパリティグループを構成するまでの期間一時的に保持するためのキャッシュ内領域の大きさは小さくてよい。

【0102】(実施例2) 本実施例では実施例1の図1に示すADC2に対し、データの属性により動的にアドレス変換を行なうかをADC2内で独自に判断する機能を付加する。本実施例ではシーケンシャルデータは、ADC2のMP1-20が独自に判断して、動的なアドレス変換を行なう領域に格納するように制御する。実施例1では、CPU1から指定されるアドレス(CPU指定アドレス)71をMP20はアドレステーブル70または動的アドレス変換テーブル90により、実際にデータの読みだしままたは書き込みを行なうSCSIドライブ12に対するSCSIドライブアドレス72に変換する。書き込み時はこの様なアドレス変換後、CPU1から送られてきたデータを一端キャッシュメモリ7に格納する。その後は、キャッシュ独自のリプレースアルゴリズムでそのデータが追い出されるまで、そのデータをキャッシュに保持する。

【0103】本実施例では書き込みデータは全て一端キ

31

ヤッシュメモリ7に格納し、一定時間キャッシュメモリ7内に保持する。このキャッシュメモリ7内に保持する時間は、ユーザがADC2に対し予め設定することができ、ADC2内のMP20により制御される。キャッシュメモリ7内に保持されたデータは、後述するように後に発行された書き込み要求のCPU指定アドレス71とでCPU指定アドレス71の比較を行ない、それぞれがシーケンシャルデータに属するかどうかを判断し、シーケンシャルデータに属するデータは動的アドレス変換を行なう領域に格納するようにMP20が独自に制御を行なう。

【0104】そこで、以下にMP1 20におけるシーケンシャルデータかどうかを判断する方法について示す。

【0105】書き込み時において、CPU1からキャッシュメモリ7内へのデータの格納後、MP1 20はCPU1から送られたCPU指定アドレス71に対し、キャッシュメモリ7内のアドレス変換用テーブルの参照を行う。図3に示すように本実施例ではアドレステーブル70の各CPU指定アドレス71に対し、アクセスフラグ84を設定する。このアクセスフラグ84は書き込み時においてCPU1が指定したCPU指定アドレス71に対し、キャッシュメモリ7内への書き込みデータの格納後、MP1 20がオン(1)とする。ユーザが予め指定した一定時間経過後、キャッシュメモリ7からSCSIドライブ12へ格納した時点でMP1 20によりアクセスフラグ84はオフ(0)とされる。

【0106】書き込み時において、CPU1から指定されたCPU指定アドレス71について、MP1 20はアドレステーブル70においてアクセスフラグ84がオン(1)になっているCPU指定アドレス71を調べ、CPU1から指定されたCPU指定アドレス71とアドレステーブル70においてアクセスフラグ84がオン(1)になっている当該CPU指定アドレス71を比較する。具体的には、ドライブ番号74とCCHHR75のシリンドアドレス(CC)を比較する。もし、前に発行され、キャッシュメモリ7内に格納されている(アクセスフラグ84がオン(1))データのCPU指定アドレス71と、後に発行された書き込み要求のCPU指定アドレス71の比較において、ドライブ番号74が一致し、しかも、CCHHR75のシリンドアドレスが一致した場合、これらの書き込み処理はシーケンシャル処理と判定する。同様に次に発行された書き込み要求についてもCPU指定アドレス71を比較する。この様にして、発行されてきた書き込み要求のデータについてMP1 20は、ユーザか予め設定しておいた一定時間の間シーケンシャル処理かどうかの判定を行なう。もし、シーケンシャルなデータの場合は、MP1 20は独自にそのシーケンシャルなデータのグループに対し、当該SCSIドライブ12の動的なアドレス変換を行なう領域に書き

10

20

30

40

50

32

込む制御を行なう。具体的には、実施例1と同様にアドレステーブル70および動的アドレステーブル90の登録を行ない、キャッシュメモリ7から当該SCSIドライブ12に対する書き込み処理をする。

【0107】(実施例3) 本実施例では図8に示すように論理グループ10単位にサブDKC11を設け、その内部に実施例1、2において示したキャッシュメモリ7内のアドレステーブル70および動的アドレス変換テーブル90を持たせたものを示す。本実施例におけるデータの処理手順の内、実施例1および2で示した処理手順と異なる部分のみを図9および10を用いて説明する。本実施例では図10に示すように実施例1、2で示したキャッシュメモリ7内のアドレステーブル70および動的アドレス変換テーブル90を各論理グループ10単位のサブDKC11内のデータアドレステーブル(DAT)30に分割する。DAT30は格納されているテーブルの形式や機能は実施例1、2と同様であるが、異なるのは書き込み又は読み出しデータを格納するメモリとは別のアドレス変換用テーブルを格納する専用メモリ内に保持される点である。ADC2内のキャッシュメモリ7内には、図5に示すようなグループアドレステーブル(GAT)23が格納されており、MP20はこのGAT23によりCPU1から指示されたCPU指定アドレス71を、そのCPU指定アドレス71が指示する場所がADU3内のどの論理グループ10かを判定する。

【0108】CPU1からの読み出し要求が転送されたときにはCPU1から指定されたCPU指定アドレス71により、MP20はGAT23で論理グループ10を確定し、MP20はこの当該論理グループ10に対し読み出し要求を発行するようにドライブインターフェース28に指示する。MP20から指示を受けたドライブインターフェース28は当該論理グループ10のサブDKC11に対し読み出し要求を発行する。サブDKC11ではマイクロプロセッサMP29がこの読み出し要求のコマンドを受け付け、DAT30を参照し、実施例1で示したMP20がアドレステーブル70および動的アドレステーブル90を用いて処理したのと同様に当該データが格納されている論理グループ10内のCPU指定アドレス71に対するSCSIドライブアドレス72およびパリティドライブアドレス73を確定する。このアドレスの確定後、実施例1でのMP20の処理と同様に、MP29は当該SCSIドライブ12に対し、読み出し要求を発行する。MP29から読み出し要求を発行されたSCSIドライブ12ではシーク、回転待ちを行ない、当該データの読み出しが可能になり次第当該データをドライブアダプタ回路34に転送し、ドライブアダプタ34はサブキャッシュメモリ32に格納する。サブキャッシュメモリ32に当該データの格納が完了した後、ドライブアダプタ34はMP29に格納報告を行ない、MP29は実施例1でのMP20と同様にDAT30内のCP

U指定アドレス71に対応した当該キャッシュフラグ82をオン(1)とし、キャッシュアドレス81にサブキャッシュ32内の格納したアドレスを登録する。後に当該キャッシュフラグ82がオンのデータに対し読み出しまたは書き込み要求が発行された場合は、サブキャッシュ32内で処理を行なう。MP3-29によるDAT30の更新が終了すると、MP3-29はADC2内のDriveIF28に対しデータ転送可能という応答を行ない、DriveIF28はこの応答を受け取ると、MP1-20に対し報告する。MP1-20はこの報告を受け取ると、キャッシュメモリ7への格納が可能なら、DriveIF28に対しサブDKC11からデータを転送するように指示する。DriveIF28ではMP1-20からの指示を受けるとサブDKC11のMP3-29に対し読み出し要求を発行する。この読み出し要求を受けたMP3-29はサブキャッシュアダプタ回路(SCA)31に対しサブキャッシュ32から当該データを読みだすように指示し、SCA31は実際にデータを読み出してDriveIF28にデータを転送する。DriveIF28がデータを受け取った後は、実施例1、2で示した処理を行なう。

【0109】一方書き込み時は読み出し時と同様に当該論理グループ10を確定し、当該論理グループ10のMP3-29に対し書き込み要求を発行する。書き込み要求を受け付け、書き込みデータをサブキャッシュ32に格納した後は、動的にアドレス変換して格納するデータについては、実施例1、2に示すように動的アドレス変換してDriveIF28に書き込み、動的アドレス変換しない場合はレベル5の処理を行なう。なお、本実施例ではサブDKC11内のサブキャッシュ32に格納されたデータにおいてPG36がパリティデータを作成する。この様にサブDKC11においてドライブ12でのアドレス管理とデータ処理を行なうことにより負荷分散が図れる。

【0110】なお、図8では各論理グループ10においてサブDKC11が付いているが、一つのサブDKC11が複数の論理グループ10を管理しても構わない。このように複数の論理グループ10に対しサブDKC11をまとめておくことによりDAT30の一括管理が行える。

【0111】なお、本実施例1、2は、磁気ディスクについて述べてきたが、光ディスク、フロッピディスクなどの記憶装置においても成立つ。

【0112】

【発明の効果】以上述べたように、本発明に従って、複数の書き込みデータとそれらから生成されたパリティデータを含むパリティグループをドライブ内の空き領域に書き込む方法をとれば、データの書き換えを従来より高速に行なえる。

【0113】さらに、本発明に従って、ある一つのパリティグループを構成する複数のデータの一部が書き換え

られたとき、書き換えられたデータは無効とし、それ以外のデータはそのまま有効なデータとしてドライブに保持するようにすれば、データの書き換え時の処理が早くなる。

【0114】さらに、本発明に従って、一つのパリティグループを構成する複数のデータの長さを上位装置より送られて来るデータの長さにすれば、いずれかのパリティグループ内のデータが部分的に無効にされたとしても、部分的に無効な状態のままでいるパリティグループは、より長いデータからパリティデータを構成した場合より少くなり、結果として、詰め替えを要する部分的に無効なパリティグループの数は少なくなる。そのため、詰め替え処理の対象となるデータは少なくなる。さらに、上位装置から転送された複数の書き込みデータを、それらからパリティグループを構成するまでの期間一時的に保持するためのキャッシュ内領域の大きさは小さくてよい。

【図面の簡単な説明】

【図1】第1の実施例の全体構成図。

【図2】第1の実施例のクラスタ内構成図。

【図3】アドレステーブルの説明図。

【図4】動的アドレス変換テーブルの説明図。

【図5】第3の実施例で用いるグループアドレステーブル(GAT)の説明図。

【図6】本発明のデータ格納方法の説明図。

【図7】本発明の詰め替え動作説明図。

【図8】第3の実施例の全体構成図。

【図9】第1の実施例のクラスタ内構成図。

【図10】図9のサブDKC内構成図。

【図11】従来のディスクアレイの説明図。

【図12】従来技術の書き込み時間の説明図。

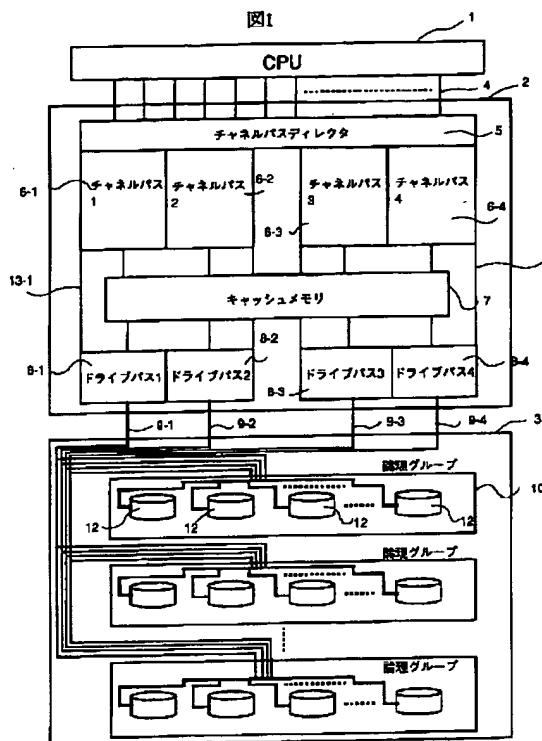
【図13】本発明のドライブの内部アドレス説明図。

【符号の説明】

2…アレイディスクコントローラ(ADC)、3…アレイディスクユニット(ADU)、4…外部インターフェースバス、5…チャネルバスディレクタ、6…チャネルバス、7…キャッシュメモリ、8…ドライブバス、9…アレイディスクユニットバス、12…ドライブ、13…クラスタ、14…ドライブ側キャッシュアダプタ(CAdp)、15…インターフェースアダプタ、16…チャネルバススイッチ、17…制御信号線、18…データ線、19…バス、20…マイクロプロセッサ1(MP1)、21…チャネルインターフェース(CHIF)回路、22…データ制御回路(DCC)、23…グループアドレス変換回路(GAT)、24…チャネル側キャッシュアダプタ(CAdp)、27…圧縮回路、28…ドライブインターフェース回路(DriveIF)、29…マイクロプロセッサ3(MP3)、30…データアドレステーブル、31…サブキャッシュアダプタ、34…ドライブアダプタ(DriveAdp)、

35…ドライブバス、36…パリティデータ生成回路、
70…アドレステーブル、71…CPU指定アドレス、
72…SCSI ドライブアドレス、73…パリティドライ
ブアドレス、74…ドライブ番号、75…CCHHR
R、76…DMポインタ、77…SCSI ドライブ番
号、78…SCSI ドライブ内アドレス、79…パリテ

【図1】



イドライブ番号、80…パリティドライブ内アドレス、
81…キャッシュアドレス、82…キャッシュフラグ、
83…ドライブフラグ、84…アクセスフラグ、85…
論理グループアドレス、90…動的アドレス変換テーブ
ル、100…グループアドレステーブル (GAT)。

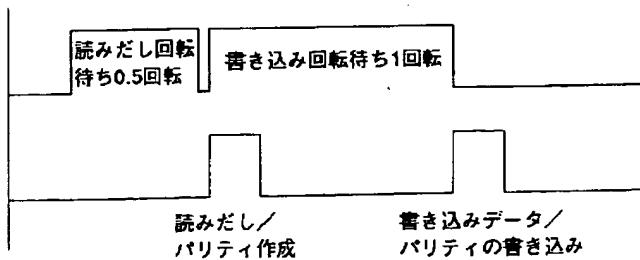
【図5】

図5

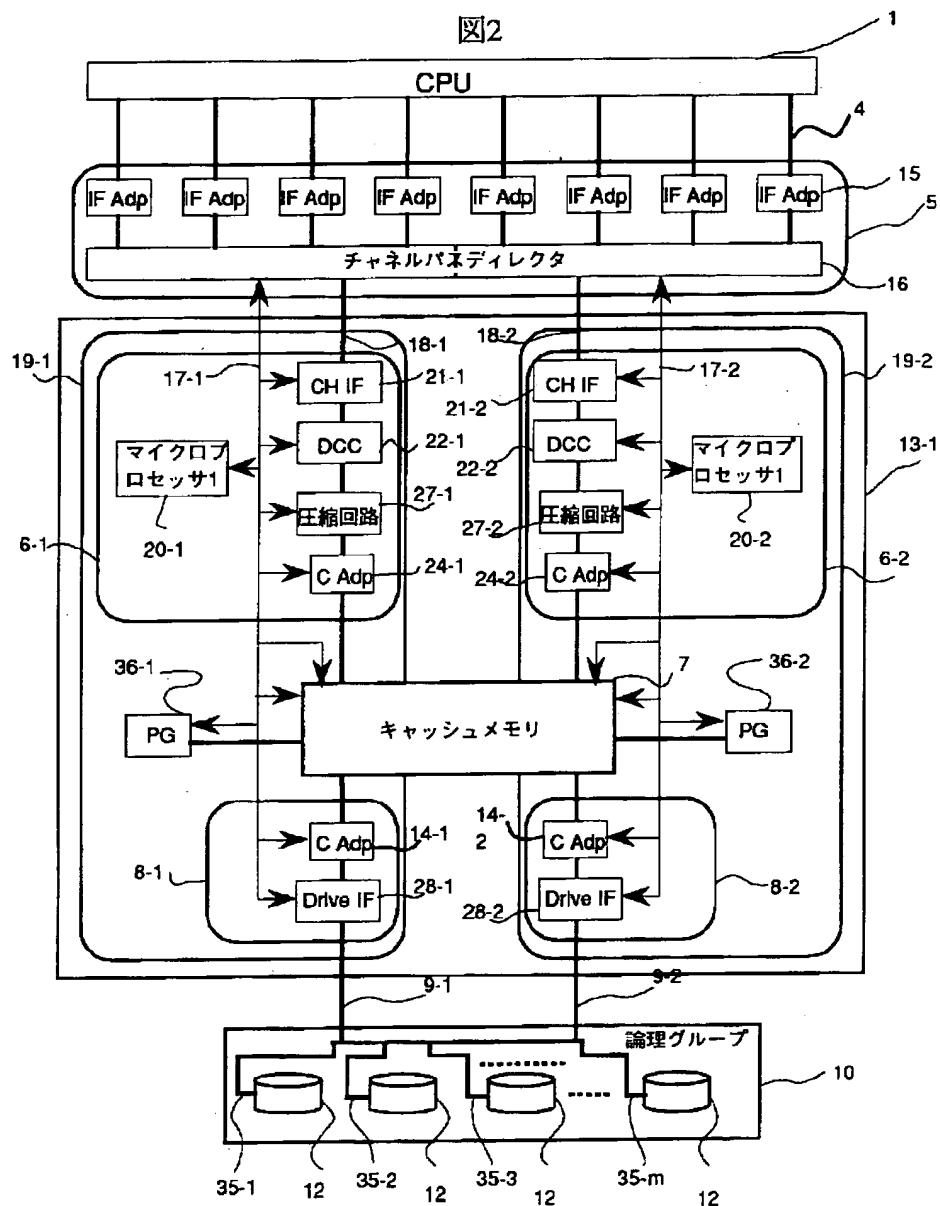
74	CPU指定アドレス	71	75	85	81	100	82
	ドライブ番号	CCHHR	論理グループアドレス	キャッシュアドレス	Cache Flag		
Drive#1	ADR 1	LADR 1	—	—	—	—	—
	ADR 2	LADR 3	—	—	—	—	—
	ADR 3	LADR 8	—	—	—	—	—
	ADR 4	LADR 2	—	—	0	—	—
Drive#2	ADR 1	LADR 1	CADR1,5	1	—	—	—
	ADR 2	LADR 5	CADR1,8	1	—	—	—
	ADR 3	LADR 4	CADR1,6	1	—	—	—
	ADR 4	—	—	—	—	—	—

【図12】

図12



【図2】

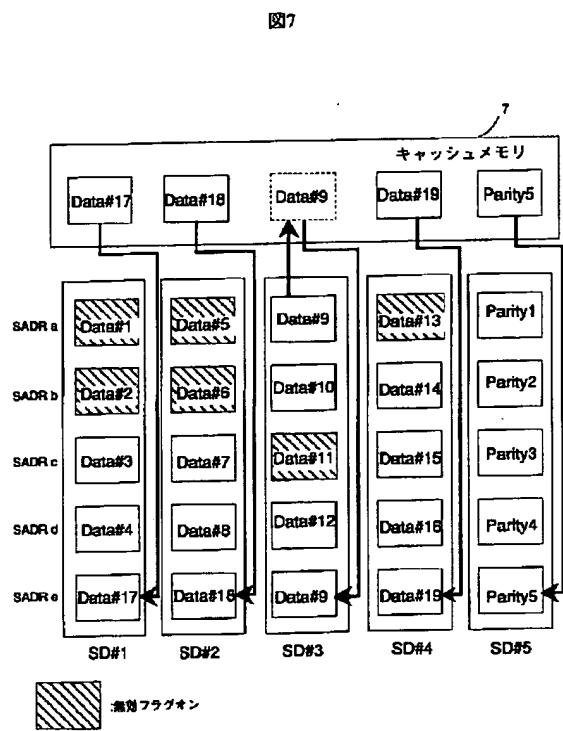


【図3】

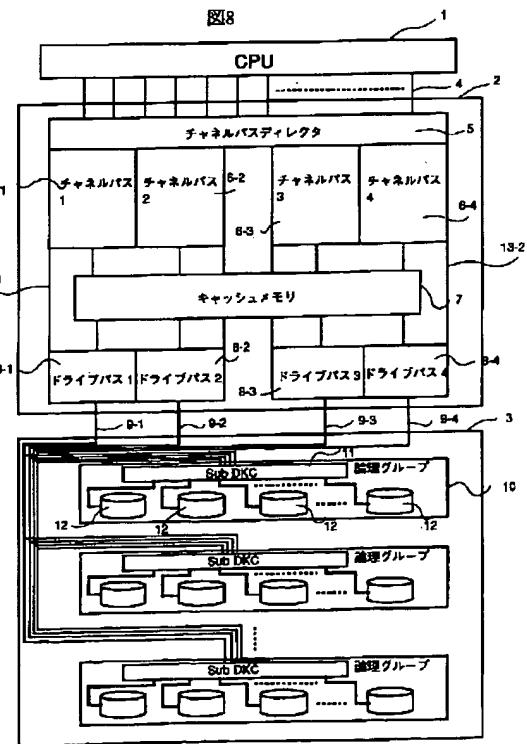
図3

CPU指定アドレス		SCSI Driveアドレス		Parity Driveアドレス		Parity内 Addr	キャッシュ アドレス	Cache Flag	アクセス フラグ
ドライブ 番号	CCHHR DM ポインタ	DM	SCSI Drive No	SCSI Addr	Parity Drive No				
Drive#1	ADR 1	DM a-2	—	—	—	—	—	—	1
	ADR 2	DM c-3	—	—	—	—	—	—	0
	ADR 3	DM d-2	—	—	—	—	—	—	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Drive#2	ADR 1	—	SD#8	SADR 1	SD#12	SADR 1	—	0	1
	ADR 2	—	SD#9	SADR 1	SD#12	SADR 1	CADR1,5	1	0
	ADR 3	—	SD#10	SADR 1	SD#12	SADR 1	CADR1,8	1	1
	ADR 4	—	SD#11	SADR 1	SD#12	SADR 1	CADR1,6	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

【図7】



【図8】



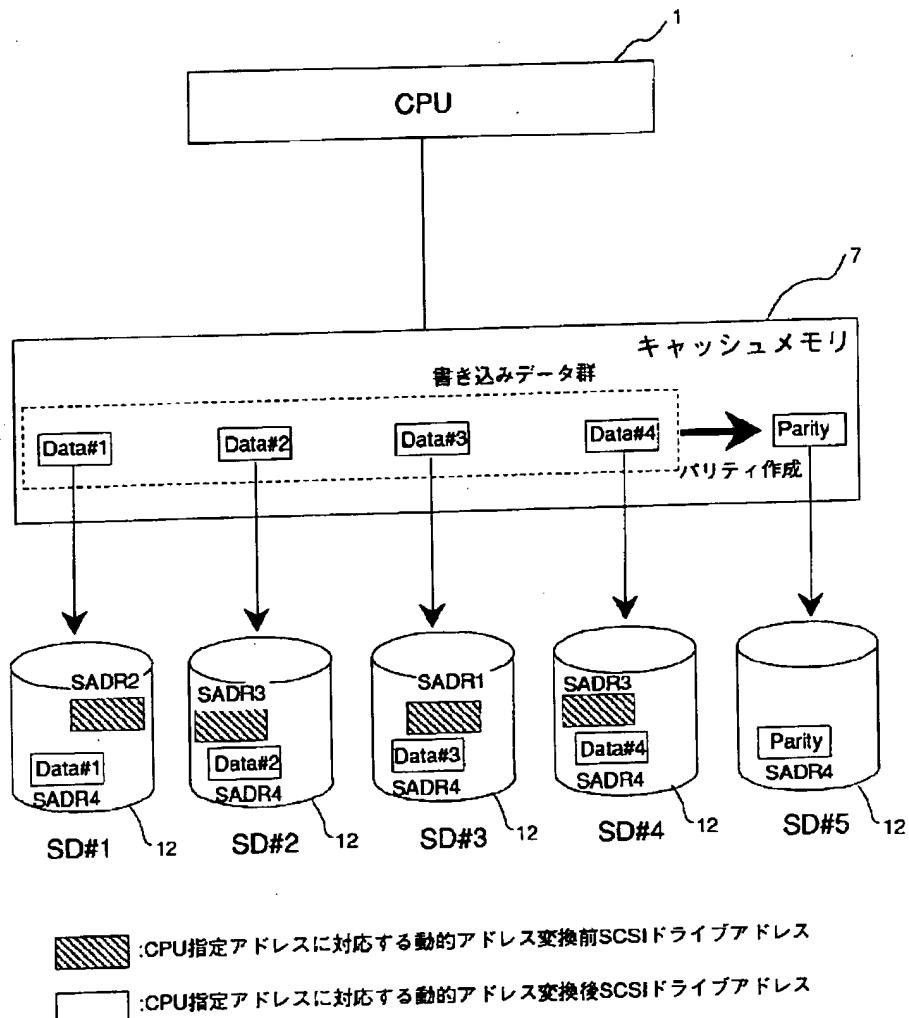
【図4】

図4

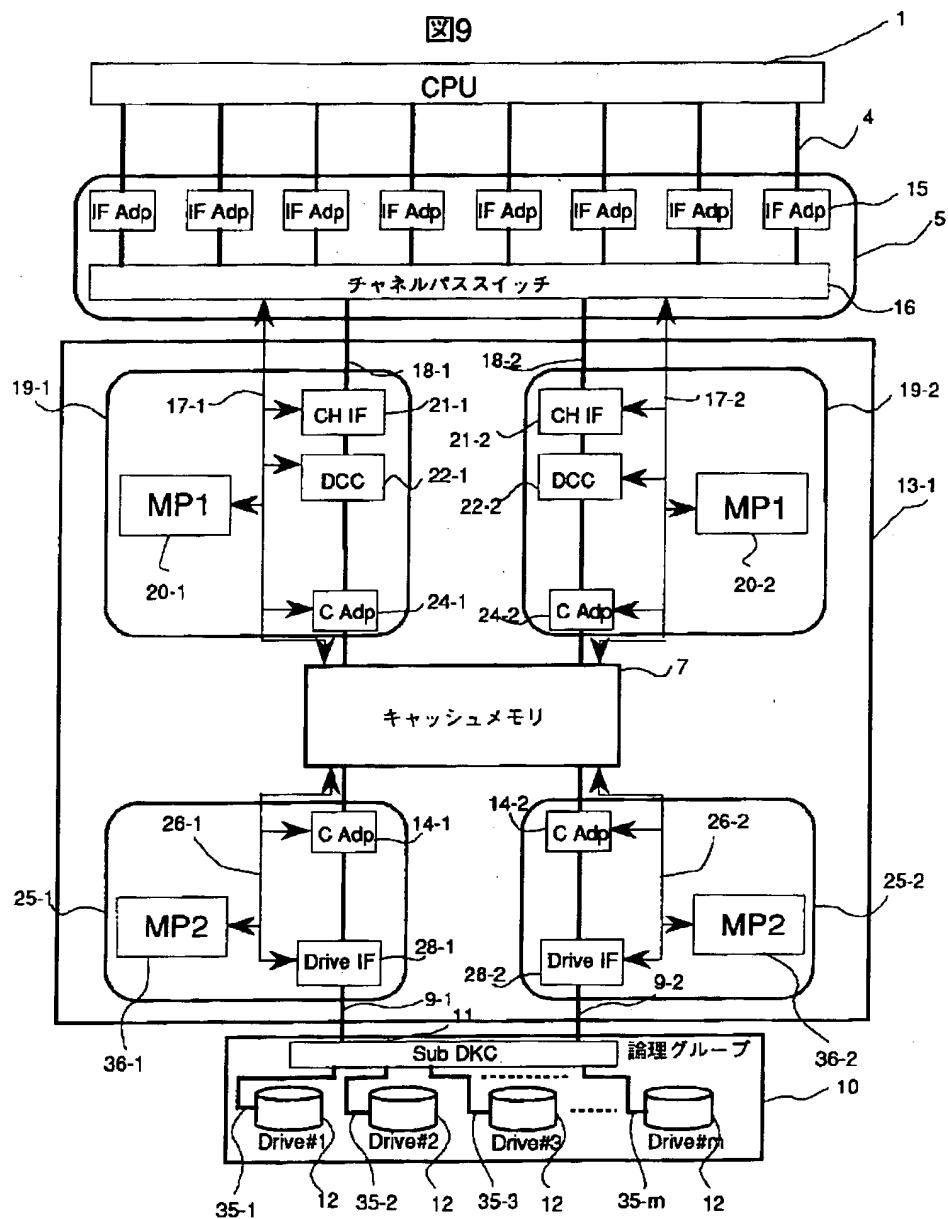
DM ポインタ	SCSI Driveアドレス		無効Flag	Parity Driveアドレス		キャッシュ アドレス	Cache Flag	Drive Flag
	SCSI Drive No	SCSI 内 Addr		Parity Drive No	Parity内 Addr			
DM a-1	SD#1	SADR 1	0	SD#5	SADR 1	CADR2.5	1	1
DM a-2		SADR 2	1	SD#5	SADR 2	CADR2.8	1	0
DM a-3		SADR 3	1	SD#5	SADR 3	—	0	0
DM a-4		SADR 4	1	SD#5	SADR 4	—	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
DM b-1	SD#2	SADR 1	0	SD#5	SADR 1	—	0	0
DM b-2		SADR 2	1	SD#5	SADR 2	—	0	0
DM b-3		SADR 3	1	SD#5	SADR 3	CADR2.3	1	0
DM b-4		SADR 4	1	SD#5	SADR 4	—	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
DM c-1	SD#3	SADR 1	1	SD#5	SADR 1	CADR2.9	0	0
DM c-2		SADR 2	1	SD#5	SADR 2	—	0	0
DM c-3		SADR 3	0	SD#5	SADR 3	—	0	1
DM c-4		SADR 4	1	SD#5	SADR 4	—	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
DM d-1	SD#4	SADR 1	1	SD#5	SADR 1	—	0	0
DM d-2		SADR 2	0	SD#5	SADR 2	CADR2.2	1	1
DM d-3		SADR 3	1	SD#5	SADR 3	CADR2.7	1	0
DM d-4		SADR 4	1	SD#5	SADR 4	—	0	1
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
DM e-1	SD#5	SADR 1	0	SD#5	SADR 1	—	0	0
DM e-2		SADR 2	0	SD#5	SADR 2	—	0	1
DM e-3		SADR 3	0	SD#5	SADR 3	—	0	0
DM e-4		SADR 4	0	SD#5	SADR 4	—	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮

【図6】

図6

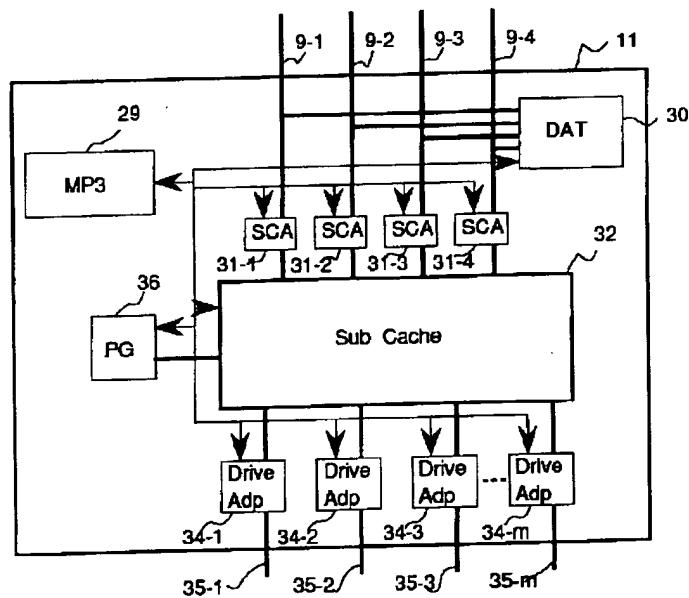


【图9】

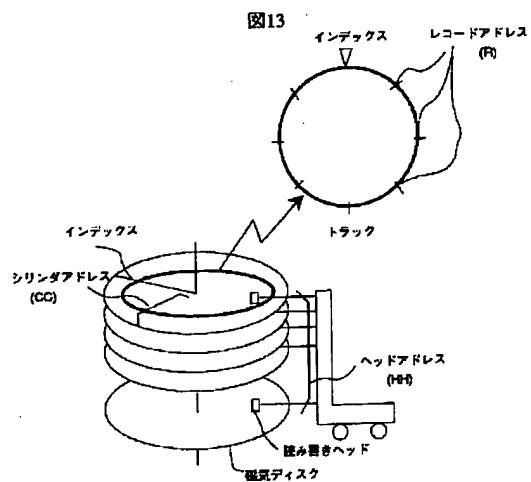


【図10】

図10



【図13】



【図11】

図11

